MODELS FOR REPEATED MEASURES OF A MULTIVARIATE RESPONSE

By

RALITZA GUEORGUIEVA

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1999

To my family

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Dr. Alan Agresti for serving as my dissertation advisor. Without his guidance, constant encouragement and valuable advice this work would not have been completed. My appreciation is extended to Drs. James Booth, Randy Carter, Malay Ghosh, and Monika Ardelt for serving on my committee and for helping me with my research. I would also like to thank all the faculty, staff, and students in the Department of Statistics for their support and friendship.

I also wish to acknowledge the members of the Perinatal Data Systems group whose constant encouragement and understanding have helped me in many ways. Special thanks go to Dr. Randy Carter, who has supported me since my first day as a graduate student.

Finally, I would like to express my gratitude to my parents, Vessela and Vladislav, for their loving care and confidence in my success, to my sister, Annie, and her fiance, Nathan, for their encouragement and continual support, and to my husband, Velizar, for his constant love and inspiration.

TABLE OF CONTENTS

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

MODELS FOR REPEATED MEASURES OF A MULTIVARIATE RESPONSE

By

Ralitza Gueorguieva

December 1999

Chairman: Alan Agresti
Major Department: Statistics

The goal of this dissertation is to propose and investigate random effects models for repeated measures situations when there are two or more response variables. The emphasis is on maximum likelihood estimation and on applications with outcomes of different types. We propose a multivariate generalized linear mixed model that can accommodate any combination of outcome variables in the exponential family. This model assumes conditional independence between the response variables given the random effects. We also consider a correlated probit model that is suitable for mixtures of binary, continuous, censored continuous, and ordinal outcomes. Although more limited in area of applicability, the correlated probit model allows for more general correlation structure between the response variables than the corresponding multivariate generalized linear mixed model.

We extend three estimation procedures from the univariate generalized linear mixed model to the multivariate generalization proposed herein. The methods are Gauss-Hermite quadrature, Monte Carlo EM algorithm, and pseudo-likelihood. Standard error approximations are considered along with parameter estimation. A simulated data example and two 'real-life' examples are used for illustration. We also

consider hypothesis testing based on quadrature and Monte Carlo approximations to the Wald, score, and likelihood ratio tests. The performance of the approximations to the test statistics is studied via a small simulation study for checking the conditional independence assumption.

We propose a Monte Carlo EM algorithm for maximum likelihood estimation in the correlated probit model. Because of the computational inefficiency of the algorithm we consider a modification based on stochastic approximations which leads to a significant decrease in the time for model fitting. To address the issue of advantages of joint over separate analyses of the response variables we design a simulation study to investigate possible efficiency gains in a multivariate analysis. Noticeable increase in the estimated standard errors is observed only in the binary response case for small number of subjects and observations per subject and for high correlation between the outcomes. We also briefly consider an identifiability issue for one of the variance components.

CHAPTER 1
INTRODUCTION

Univariate repeated measures occur when one response variable is observed at several occasions for each subject. Hereafter subject refers to any unit on which a measurement is taken, while occasion corresponds to time or to a specific condition. If more than one response is observed at each occasion, multivariate repeated measures are available. Univariate and multivariate repeated measures are very common in biomedical applications, for example when one or more variables are measured on each patient at a number of hospital visits, or when a number of questions are asked at a series of interviews. But the occasions do not necessarily refer to different times. For instance dependent responses can be measured on litter mates, on members of the same family or at different places on a subject's body. Difficulties in analyzing repeated measures arise because of correlations usually present between observations on the same subject. Statistical methods and estimation techniques are well developed for repeated measures on a univariate normal variable, and lately much research has been dedicated to repeated observations on a binary variable and more generally on variables with distributions in the exponential family. Zeger and Liang (1992) provide an overview of methods for longitudinal data and the books of Lindsey (1993), Diggle, Liang and Zeger (1994), and Fahrmeir and Tutz (1994) cover many details. Pendergast et al. (1996) present a comprehensive survey of models for correlated binary outcomes, including longitudinal data.

However, relatively little attention is concentrated on repeated measures of a multivariate response. General models for this situation are necessarily complex as two

types of correlations must be taken into account: correlations between measurements on different variables at each occasion and correlations between measurements at different occasions. Reinsel (1982, 1984), Lundbye-Christensen (1991), Matsuyama and Ohashi (1997), and Heitjan and Sharma (1997) consider models for normally distributed responses. Lefkopoulou, Moore and Ryan (1989), Liang and Zeger (1989), and Agresti (1997) propose models for multivariate binary data; Catalano and Ryan (1992), Fitzmaurice and Laird (1995), and Regan and Catalano (1999) introduce models for clustered bivariate discrete and continuous outcomes. Catalano (1994) considers an extension to ordinal data of the Catalano and Ryan model. Rochon (1996) demonstrates how generalized estimating equations can be used to fit extended marginal models for bivariate repeated measures of discrete or continuous outcomes. Rochon's approach is very general and allows for a large class of response distributions. However, in many cases, especially when subject-specific inference is of primary interest, marginal models are not appropriate as they may lead to attenuation of the estimates of the regression parameters (Zeger, Liang, and Albert, 1988).

Sammel, Ryan, and Legler (1997) analyze mixtures of discrete and continuous responses in the exponential family using latent variables models. Their approach is based on numerical or stochastic approximations to maximum-likelihood and allows for subject-specific inference. Blackwell and Catalano (1999a, 1999b) consider extensions for ordinal data and for repeated measures of ordinal responses.

The Generalized Linear Mixed Model (GLMM) forms a very general class of subject-specific models for discrete and continuous responses in the exponential family and is used for univariate repeated measures (Fahrmeir and Tutz, 1994). In the current dissertation we demonstrate how the GLMM approach can be extended for multivariate repeated measures by assuming separate random effects for each outcome variable. This is in contrast to the Sammel et al. approach in which common underlying latent variables are assumed. We also consider a more general correlated

probit model than the appropriate GLMM for the special case of clustered binary and continuous data.

The introduction to this dissertation contains an overview of approaches for modelling of univariate repeated measures (Section 1.1), existing models for multivariate repeated measures (Section 1.2), and simultaneous modeling of different types of responses (Section 1.3). The chapter concludes with an outline of the dissertation (Section 1.4).

1.1 Models for Univariate Repeated Measures

1.1.1 General Linear Models

Historically, the first models for repeated measures data to be considered are general linear models with correlated normally distributed errors (see Ware, 1982, for review). The univariate respresentation of such models is as follows:

Suppose each subject i is observed on J occasions and denote the column vector of responses by $\mathbf{y}_i$. Assume that $\mathbf{y}_i$ arises from the linear model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \tag{1.1}$$

where $\mathbf{X}_i$ is a $J \times p$ model matrix for the i^{th} individual, $\boldsymbol{\beta}$ is a $p \times 1$ unknown parameter vector and $\boldsymbol{\epsilon}_i$ is a $J \times 1$ error vector with a multivariate normal distribution with mean $\mathbf{0}$ and arbitrary positive definite covariance matrix $\boldsymbol{\Sigma}$: $N_J(\mathbf{0}, \boldsymbol{\Sigma})$. $\boldsymbol{\Sigma}$ can take certain special forms. The case $\boldsymbol{\Sigma} = \mathbf{I}$ corresponds to the usual linear model for cross-sectional data. The equicorrelation structure ($\boldsymbol{\Sigma} = \sigma^2(\rho\mathbf{I} + (1-\rho)\mathbf{J})$, where $\mathbf{J}$ is a matrix of ones, $0 < \rho \leq 1$ and $\sigma^2 > 0$) is appropriate when the repeated measures are on subjects within a cluster, for example when a certain characteristic is observed on each of a number of litter mates. The autoregressive structure for $\boldsymbol{\Sigma} = ((\sigma_{ij}))$ ($\sigma_{ij} = \rho^{|i-j|}$) is one of the most popular structures when the observations are over equally-spaced time periods.

Usually of primary interest in general linear models is the estimation of regression parameters while recognizing the likely correlation structure in the data. To achieve this, one either assumes an explicit parametric model for the covariance structure, or uses methods of inference that are robust to misspecification of the covariance structure. Weighted least squares, maximum likelihood and restricted maximum likelihood are the most popular estimation methods for general linear models.

1.1.2 Generalized Linear Models

Generalized linear models (GLM) are natural extensions of classical linear models allowing for a larger class of response distributions. Their specification consists of three parts (McCullagh and Nelder, 1989, pp.27-30): a random component, a systematic component and a link function.

1. The random component is the probability distribution for the elements of the response vector. The y_i's, $i = 1, ...n$, are assumed to be independent with a distribution in the exponential family

$$f(y_i; \theta_i, \phi_i) = exp\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i)\} \tag{1.2}$$

for some specified functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. Usually

$$a(\phi_i) = \frac{\phi}{w_i},$$

where ϕ is called a dispersion parameter and w_i are known weights. The mean μ_i and the variance function $\nu(\mu_i)$ completely specify a member of the exponential family because $\mu_i = b'(\theta_i)$ and $\nu(\mu_i) = b''(\theta_i)a(\phi_i)$. Important exponential family distributions are the normal, the binomial, the Poisson and the gamma distributions.

2. The systematic component is a linear function of the covariates

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta},$$

where η_i is commonly called a linear predictor.

3. The link function $g(\cdot)$ is a monotonic differentiable function which relates the expected value of the response distribution μ_i to the linear predictor η_i:

$$\eta_i = g(\mu_i)$$

When the response distribution is normal and the link is the identity function $g(\mu) = \mu$, the GLM reduces to the usual linear regression model. For each member of the exponential family there is a special link function, called the canonical link function, which simplifies model fitting. For that link function $\theta_i = \eta_i$. Maximum likelihood estimates in GLM are obtained using iterative reweighted least squares.

Just as modifications of linear models are used for analyzing Gaussian repeated measures, modifications of GLM can handle discrete and continuous outcomes. Extensions to GLM include marginal, random effects and transition models (Zeger and Liang, 1992). Hereafter we will use y_{ij} to denote the response at the j^{th} occasion for the i^{th} subject. The ranges of the subscripts will be $i = 1, ...n$, and $j = 1, ...J$ for balanced and $j = 1, ...n_i$ for unbalanced data.

1.1.3 Marginal Models

Marginal models are designed to permit separate modeling of the regression of the response on the predictors, and of the association among repeated observations for each individual. The models are defined by specifying expressions for the marginal mean and the marginal variance-covariance matrix of the response:

1. The marginal mean $\mu_{ij} = E(y_{ij})$ is related to the predictors by a known link function $g(\cdot)$

$$g(\mu_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta}$$

2. The marginal variance is a function of the marginal mean

$$Var(y_{ij}) = V(\mu_{ij})\phi$$

and the marginal covariance is a function of the marginal means and of additional parameters $\boldsymbol{\delta}$

$$Cov(y_{ij_1}, y_{ij_2}) = c(\mu_{ij_1}, \mu_{ij_2}; \boldsymbol{\delta}).$$

Notice that if the correlation is ignored and the variance function is chosen to correspond to an exponential family distribution, the marginal model reduces to GLM for independent data. But the variance function can be more general, and hence even if the responses are uncorrelated this model is more general than the corresponding GLM. Because only the first two moments are specified for the joint distribution of the response, additional assumptions are needed for likelihood inferences. Alternatively, the Generalized Estimating Equations (GEE) method can be used (Liang and Zeger (1986), Zeger, Liang and Albert (1988)) as briefly summarized here.

Let $\mathbf{y}_i = (y_{i1}, ...y_{i,n_i})^T$, $\boldsymbol{\mu}_i = (\mu_{i1}, ...\mu_{i,n_i})^T$, $\mathbf{A}_i = diag\{V(\mu_{i1}), ..., V(\mu_{i,n_i})\}$ and let $\mathbf{R}_i(\boldsymbol{\delta})$ be a 'working' correlation matrix for the i^{th} subject. The latter means than $\mathbf{R}_i(\boldsymbol{\delta})$ is completely specified up to a parameter vector $\boldsymbol{\delta}$ and may or may not be the true correlation matrix. The regression parameters $\boldsymbol{\beta}$ are then estimated by solving

$$\mathbf{S}_{\boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i^T}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(\boldsymbol{\delta})(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0},$$

where $\mathbf{V}_i(\boldsymbol{\delta}) = \mathbf{A}_i^{\frac{1}{2}} \mathbf{R}_i(\boldsymbol{\delta}) \mathbf{A}_i^{\frac{1}{2}}$. Liang and Zeger (1986) show that if the mean function is correctly specified, $\hat{\boldsymbol{\beta}}$, the solution to the above equation, is consistent and asymptotically normal as the number of subjects goes to infinity. They also propose a robust variance estimate, which is also consistent even when the variance-covariance structure is misspecified. Hence, the GEE approach is appropriate when the regression relationship, and not the correlation structure of the data, is of primary interest.

1.1.4 Random Effects Models

An important feature of the marginal models is that the regression coefficients have the same interpretation as coefficients from a cross-sectional analysis. These models are preferred when the effects of explanatory variables on the average response within a population are of primary interest. However, when it is of interest to describe how the response for a particular individual changes as a result of a change in the covariates, a more pertinent approach is to consider random (mixed) effects models.

Random effects models assume that the correlation among repeated responses arises because there is a natural heterogeneity across individuals and that this heterogeneity can be represented as a probability distribution. More precisely,

1. The conditional distribution of the response y_{ij} given a subject-specific vector of random effects $\mathbf{b}_i$ satisfies a GLM with a linear predictor $\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i$, where $\mathbf{z}_{ij}$ in general is a subset of $\mathbf{x}_{ij}$.

2. The responses on the same subject $y_{i1}, y_{i2},..., y_{in_i}$ are conditionally independent given $\mathbf{b}_i$.

3. $\mathbf{b}_i$ has certain distribution $F(\cdot;\boldsymbol{\delta})$ with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ depending on a parameter vector $\boldsymbol{\delta}$.

Such models with normal distribution for the random effects are considered in greater detail in Section 2.1. In contrast to marginal models, the regression coefficients in random effects models have subject-specific interpretations. To better illustrate that difference let us consider a particular example.

Let the response y_{ij} be binary and the subscript j refer to time. Consider the marginal model

$$logit(\mu_{ij}) = \beta_0 + \beta_1 j,$$

where $\mu_{ij} = E(y_{ij})$, and the random effects model

$$logit(\mu_{ij}^c) = \beta_0^c + \beta_1^c j + b_i,\ b_i \sim\ i.i.d.\ N(0, \sigma^2),$$

where $\mu_{ij}^c = E(y_{ij}|b_i)$. Then β_1^c is the log-odds ratio for a positive response at time $j+1$ relative to time j for any subject i, while β_1 is the population-averaged log-odds ratio. That is β_1 describes the change in log-odds for a positive response from time j to time $j+1$ for the population as a whole. In general these two interpretations are different but in some special cases such as identity link functions subject-specific and population-averaged interpretations coincide. More discussion on the connection between marginal and random effects models follows in Chapter 2.

The presence of random effects enables the pooling of information across different subjects to result in better subject-specific as opposed to population-averaged inference, but complicates the estimation problem considerably. To obtain the likelihood function, one has to integrate out the random effects, which, except for a few special cases, cannot be performed analytically. If the random effects are nuisance parameters, conditional likelihood estimates for the fixed effects may be easy to obtain for canonical link functions. This is accomplished by conditioning on the sufficient statistics for the unknown nuisance parameters and then maximizing the conditional likelihood.

When the dimension of the integral is not high, numerical methods such as Gaussian quadrature work well for normally distributed random effects (Fahrmeir and Tutz, 1994, pp.357-362; Liu and Pierce, 1994). A variety of other methods have recently been proposed to handle more difficult cases. These include the approximate maximum likelihood estimates proposed by Schall (1991), the penalized quasi-likelihood approach of Breslow and Clayton (1993), the hierarchical maximum likelihood of Lee and Nelder (1996), the Gibbs sampling approach of Zeger and Karim (1991), the EM algorithm approach for GLMM of Booth and Hobert (1999) and of McCulloch (1997) and others.

1.1.5 Transition Models

A third group of models for dealing with longitudinal data consists of transition models. Regression Markov chain models for repeated measures data have been considered by Zeger et al. (1985) and Zeger and Qaqish (1988). This approach involves modeling the conditional expectation of the response at each occasion given past outcomes. Specifically,

1. The conditional expectation of the response $\mu_{ij}^c = E(y_{ij}|y_{i,j-1},...y_{i1})$ depends on previous responses and current covariates as follows

$$g(\mu_{ij}^c) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \sum_{k=1}^{j} \gamma_k f_k(y_{i,j-1},...y_{i1}),$$

where $\{f_k(\cdot)\}, k = 1,...j$ are known functions.

2. The conditional variance of y_{ij} is a function of the conditional mean

$$Var(y_{ij}|y_{i,j-1},...y_{i1}) = V(\mu_{ij}^c)\phi,$$

where V is a known function.

Transition models combine the assumptions about the dependence of the response on the explanatory variables and the correlation among repeated observations into a single equation. Conditional maximum likelihood and GEE have both been used for estimating parameters.

Extensions of GLM are not the only methods for analyzing repeated measures data (see for example Vonesh (1992) for an overview of non-linear models for longitudinal data), but as the proposed models in this dissertation are based on such extensions, our discussion in the following sections will be restricted to GLM types of models.

1.2 Models for Multivariate Repeated Measures

In contrast to univariate longitudinal data, very few models have been discussed that specifically deal with multivariate repeated measures. These models are now briefly discussed, starting with normal theory linear models and proceeding with models for discrete outcomes.

A review of general linear models for the analysis of longitudinal studies is provided by Ware (1985). The general multivariate model is defined as in (1.1) but we assume that the $J = KL$ repeated measures on each subject are made on K normally distributed variables rather than on only one normally distributed response. Hence, the general multivariate model with unspecified covariance structure can be directly applied to multivariate repeated measures data. However, the number of parameters to be estimated increases quickly as the number of occasions and/or variables increases and the estimation may become quite burdensome. Special types of correlation structures such as bivariate autoregressive can be specified directly as proposed by Galecki (1994). This multivariate linear model is also not well suited for unbalanced or incomplete data.

More parsimonious covariance structures are achieved in random effects models. The linear mixed effects model is defined in two stages. At stage 1 we assume

$$\mathbf{y}_i|\mathbf{b}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{\epsilon}_i$ is distributed $N_{n_i}(\mathbf{0}, \sigma^2\mathbf{I})$, $\mathbf{X}_i$ and $\mathbf{Z}_i$ are $n_i \times p$ and $n_i \times q$ model matrices for the fixed and the random effects respectively, and $\mathbf{b}_i$ is a $q \times 1$ random effects vector. At stage 2, $\mathbf{b}_i \sim N_q(\mathbf{0}, \boldsymbol{\Sigma})$ independently of $\boldsymbol{\epsilon}_i$. This corresponds to a special variance-covariance structure for the i^{th} subject: $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{I} + \mathbf{Z}_i\boldsymbol{\Sigma}\mathbf{Z}_i^T$.

Reinsel (1982) generalized this linear mixed effects model and showed that for balanced multivariate repeated measures models with random effects structure, closed form solutions exist for both maximum likelihood (ML) and restricted maximum

likelihood (REML) estimates of mean and covariance parameters. Matsuyama and Ohashi (1997) considered bivariate response mixed effects models that can handle missing data. Choosing a Bayesian viewpoint, they used the Gibbs sampler to estimate the parameters.

The model of Reinsel is overly restrictive in some cases as he prescribes the same growth pattern over all response variables for all individuals. In contrast, Heitjan and Sharma (1997) considered a model for repeated series longitudinal data, where each unit could yield multiple series of the same variable. The error term they used was a sum of a random subject effect and a vector autoregressive process, thus accounting for subject heterogeneity and time dependence in an additive fashion. A straightforward extension for multiple series of observations on distinct variables is possible.

All models discussed so far in this section are appropriate for continuous response variables that can be assumed to be normally distributed. More generally GEE marginal models can easily accommodate multivariate repeated measures if all response variables have the same discrete or continuous distribution. We now restate the GEE marginal model from Section 1.1 in matrix notation to simplify the multivariate extension. We also consider balanced data. Let $\mathbf{y}_i$ represent the vector of observations for the i^{th} individual $(i = 1, 2, ...n)$ and let $\boldsymbol{\mu}_i = E(\mathbf{y}_i)$ be the marginal mean vector. We assume

$$g(\boldsymbol{\mu}_i) = \mathbf{X}_i\boldsymbol{\beta},$$

where $\mathbf{X}_i$ is the model matrix for the i^{th} individual, $\boldsymbol{\beta}$ is an unknown parameter vector and $g(\cdot)$ is a known link function, applied componentwise to $\boldsymbol{\mu}_i$. Also $Var(y_{ij}) = \phi V(\mu_{ij})$, where V is a known variance function and ϕ is a dispersion parameter. Letting $\mathbf{A}_i = diag[V(\mu_{i1}), ...V(\mu_{iJ})]$ and assuming a working correlation matrix $\mathbf{R}_i(\boldsymbol{\delta})$ for $\mathbf{y}_i$, the 'working' covariance matrix for $\mathbf{y}_i$ is given by

$$\mathbf{V}_i = \phi(\mathbf{A}_i)^{\frac{1}{2}}\mathbf{R}_i(\boldsymbol{\delta})(\mathbf{A}_i)^{\frac{1}{2}}.$$

If $\mathbf{R}_i(\boldsymbol{\delta}) = \mathbf{R}(\boldsymbol{\delta})$ is the true correlation matrix, $\mathbf{V}_i$ is the true covariance matrix of $\mathbf{y}_i$. The J responses for each subject are usually repeated measures on the same variable, but as in the normal case, they can be repeated observations on two or more outcome variables as long as they have the same distribution. The only difference with univariate repeated measures is in the specification of the covariance matrix. The estimates of the regression parameters $\boldsymbol{\beta}$ will be consistent provided that the model for the marginal mean structure is specified correctly, but for better efficiency, the working correlation matrix should be close to the true correlation matrix. Several correlation structures for multivariate repeated measures are discussed by Rochon (1996).

As a further extension, Rochon (1996) proposed a model for bivariate repeated measures that could accommodate both continuous and discrete outcomes. He used GEE models as the one described above, to relate each set of repeated measures to important explanatory variables and then applied seemingly unrelated regression (SUR, Zellner, 1962) methodology to combine the pair of GEE models into an overall analysis framework. If the response vector for the i^{th} subject is denoted by $\mathbf{y}_i = (\mathbf{y}_i^{(1)^T}, \mathbf{y}_i^{(2)^T})^T$ and all relevant quantities for the first and second response are superscribed by 1 and 2 respectively, the SUR model may be written as

$$g(\boldsymbol{\mu}_i) = \mathbf{X}_i\boldsymbol{\beta},$$

where

$$\boldsymbol{\mu}_i = \begin{bmatrix} \boldsymbol{\mu}_i^{(1)} \\ \boldsymbol{\mu}_i^{(2)} \end{bmatrix}, \ \mathbf{X}_i = \begin{bmatrix} \mathbf{X}_i^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_i^{(2)} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}^{(1)} \\ \boldsymbol{\beta}^{(2)} \end{bmatrix},$$

and $g(\cdot)$ is a compound function consisting of $g^{(1)}(\cdot)$ and $g^{(2)}(\cdot)$. The joint covariance matrix among the sets of repeated measures may be written as

$$\mathbf{V}_i = \begin{bmatrix} \phi_1\mathbf{I}_J & \mathbf{0} \\ \mathbf{0} & \phi_2\mathbf{I}_J \end{bmatrix}^{1/2} \begin{bmatrix} \mathbf{A}_i^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_i^{(2)} \end{bmatrix}^{1/2} \mathbf{R}(\boldsymbol{\delta}) \begin{bmatrix} \mathbf{A}_i^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_i^{(2)} \end{bmatrix}^{1/2} \begin{bmatrix} \phi_1\mathbf{I}_J & \mathbf{0} \\ \mathbf{0} & \phi_2\mathbf{I}_J \end{bmatrix}^{1/2},$$

where $\mathbf{R}(\boldsymbol{\delta})$ is the working correlation matrix among the two sets of repeated measures for each subject, and each of its elements is a function of the vector parameter δ.

$$\mathbf{R}(\delta) = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{12}^T & \mathbf{R}_{22} \end{bmatrix}$$

The suggested techniques may be extended to multiple outcome measures. Rochon's approach provides a great deal of flexibility in modeling the effects of both within-subject and between-subject covariates on discrete and continuous outcomes and is appropriate when the effects of covariates on the marginal distributions of the response are of interest. If subject-specific inference is preferred, however, a transitional or random effects models should be considered.

Liang and Zeger (1989) suggest a class of Markov chain logistic regression models for multivariate binary time series. Their approach is to model the conditional distribution of each component of the multivariate binary time series given the others. They use 'pseudolikelihood' estimation methods to reduce the computational burden associated with maximum likelihood estimation.

Liang and Zeger's transitional model is useful when the association among variables at one time is of interest, or when the purpose is to identify temporal relationship among the variables, adjusting for covariates. However, one must use caution when interpreting the estimated parameters because the regression parameter $\boldsymbol{\beta}$ in the model has log-odds ratio interpretation conditional on the past and on the other outcomes at time t. Hence, if a covariate influences more than one component of the outcome vector or past outcomes, which is frequently the case, its regression coefficient will capture only that part of its influence that cannot be explained by the other outcomes it is also affecting. Another problem with the model is that in the fitting process all the information on the first q observations is ignored.

A random effects approach for dealing with multivariate longitudinal data is discussed by Agresti (1997) who develops a multivariate extension of the Rasch model for

repeated measures of a multivariate binary response. One disadvantage of the multivariate Rasch model shared by many random effects models, is that it can not model negative covariance structure among repeated observations on the same variable. Although usually measurements on the same variable within a subject are positively correlated, there are cases when the correlation is negative. One such example is the infection data, first considered by Haber (1986). Frequencies of infection profiles of a sample of 263 individuals for four influenza outbreaks over four consecutive winters in Michigan were recorded. The first and fourth outbreaks are known to be caused by the same virus type and because contracting influenza during the first outbreak provides an immunity against a subsequent outbreak, a subject's response for these two outbreaks is negatively correlated. Coull (1997) analyzed these data using the multivariate binomial-logit normal model. As it is a special case of the models that we propose later in this dissertation, we define it here.

Let $\mathbf{y} = (y_1, ...y_I)^T$ given $\boldsymbol{\pi} = (\pi_1, ...\pi_I)^T$ be a random vector of independent binomial components with number of trials $(n_1, ...n_I)^T$. Also let $logit(\boldsymbol{\pi})$ be $N_I(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $\boldsymbol{\pi}$ has multivariate logistic-normal distribution and unconditionally $\mathbf{y}$ has a multivariate binomial logit-normal mixture distribution. If the I observations correspond to measurements on K variables at L occasions, this model can be used for analyzing multivariate repeated measures data. The mean of the multivariate random effects distribution can be assumed to be a function of covariates $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and several groups of subjects with the same design matrices $\mathbf{X}_s$, $s = 1, ...S$, can be considered.

The multivariate binomial-logit normal model can be regarded as an analog for binary data of the multivariate Poisson-log normal model of Aitchison and Ho (1989) for count data. Aitchison and Ho assume that $\mathbf{y} = (y_1, ...y_I)^T$ given $\boldsymbol{\theta} = (\theta_1, ...\theta_I)^T$ are independent Poisson with mean vector $\boldsymbol{\theta}$, and that $log(\boldsymbol{\theta}) = (log(\theta_1), ...log(\theta_I))^T$ are $N_I(\mu, \boldsymbol{\Sigma})$. Then $\boldsymbol{\theta}$ has multivariate log-normal distribution. Like the multivariate

logit-normal model, the multivariate Poisson-log normal model can be used to model negative correlations and can be extended to incorporate covariates.

Chan and Kuk (1997) also consider random effects models for binary repeated measures but assume an underlying threshold model with normal errors and random effects, and use the probit link. They estimate the parameters via a Monte Carlo EM algorithm, regarding the observations from the underlying continuous model as the complete data. Their approach will be discussed in more detail in Chapter 5 where an extension of their model will be considered.

1.3 Simultaneous Modelling of Responses of Different Types

Difficulties in joint modelling of responses of different types arise because of the need to specify a multivariate joint distribution for the outcome variables. Most research so far has concentrated on simultaneous analysis of binary and continuous responses.

Olkin and Tate (1961) introduced a location model for discrete and continuous outcomes. It is based on a multinomial model for the discrete outcomes and a multivariate Gaussian model for the continuous outcomes conditional on the discrete outcomes. Fitzmaurice and Laird (1995) discussed a generalization of this model which turns out to be a special case of the partly exponential model introduced by Zhao, Prentice and Self (1992). Partly exponential models for the regression analysis of multivariate (discrete, continuous or mixed) response data are parametrized in terms of the response mean and a general shape parameter. They encompass generalized linear models as well as certain multivariate distributions. A fully parametric approach to estimation leads to asymptotically independent maximum likelihood estimates of the mean and the shape parameters. The score equations for the mean parameters are essentially the same as in GEE. The authors point out two major drawbacks to their approach: one is the computational complexity of full maximum

likelihood estimation of the mean and the shape parameters together, another one is the need to specify the shape function for the response. The latter will be especially hard if the response vector is a mixture of discrete and continuous outcomes. Zhao, Prentice and Self conclude that partly exponential models are mainly of theoretical interest and can be used to evaluate properties of other mean estimation procedures.

Cox and Wermuth (1992) compared a number of special models for the joint distribution of qualitative (binary) and quantitative variables. The joint distribution for all bivariate models is based on the marginal distribution of one of the components and a conditional distribution for the other component given the first one. A key distinction is between models in which for each binary outcome (A), the quantitative response (Y) is assumed to be normally distributed and models in which the marginal distribution of Y is normal. Typically, simplicity in the marginal distribution of Y corresponds to fairly complicated conditional distribution of Y and vice versa but normality often holds at least approximately. Estimation procedures differ from model to model but essentially the same tests of independence of the two components of the response can be derived. This, however, is not true if trivariate distributions are considered with two binary and one continuous, or with two continuous and one binary component. In that case several different hypotheses of independence and conditional independence can be considered and depending on the model sometimes they may not be tested unless a stronger hypothesis of independence is assumed.

Catalano and Ryan (1992), Fitzmaurice and Laird (1995) and Regan and Catalano (1999) considered mixed models for a bivariate response consisting of a binary and of a quantitative variable. Catalano and Ryan (1992) and Regan and Catalano (1999) treated the binary variable as a dichotomized continuous latent trait, which had a joint bivariate normal distribution with the other continuous response. Catalano and Ryan (1992) then parametrized the model so that the joint distribution was a product of a standard random effects model for the continuous variable and a correlated probit

model for the discrete variable. Estimation for the parameters of the two models is performed using quasi-likelihood techniques. Catalano (1994) extended the Catalano and Ryan procedure to ordinal instead of binary data.

Regan and Catalano (1999) used exact maximum likelihood for estimation, which is computationally feasible because of the equicorrelation assumption between and within the binary and continuous outcomes. The maximum likelihood methodology used by the authors is an extension to the procedure suggested by Ochi and Prentice (1984) for binary data.

Fitzmaurice and Laird (1995) assumed a logit model for the binary response and a conditional Gaussian model for the continuous response. Unlike Catalano and Ryan's model all regression parameters have marginal interpretations and the estimates of the regression parameters (based on ML or GEE) are robust to misspecification of the association between the binary and the continuous responses.

Sammel, Ryan and Legler (1997) developed models for mixtures of outcomes in the exponential family. They assumed that all responses are manifestations of one or more common latent variables and that conditional independence between the outcomes given the value of the latent trait held. This allows one to use the EM algorithm with some numerical or stochastic approximations at each E-step. Blackwell and Catalano (1999) extended the Sammel, Ryan and Legler methodology to longitudinal ordinal data by assuming correlated latent variables at each time point. For simplicity of analysis each outcome is assumed to depend only on one latent variable and the latent variables are assumed to be independent at each time point.

1.4 Format of Dissertation

The purpose of this dissertation is to propose and investigate random effects models for repeated measures situations when there are two or more response variables.

Of special interest is the case when the response varibles are of different types. The dissertation is organized as follows.

In Chapter 2 we propose a multivariate generalized linear mixed model which can accomodate any combination of responses in the exponential family. We first describe the usual generalized linear mixed model (GLMM) and then define its extension. The relationship between marginal and conditional moments in the proposed model is briefly discussed and two motivating examples are presented. The key assumption of conditional independence is outlined.

Chapter 3 concentrates on maximum likelihood model fitting methods for the proposed model. Gauss-Hermite quadrature, a Monte Carlo EM algorithm and a pseudo-likelihood approach are extended from the univariate to the multivariate generalized linear mixed model. Standard error approximation is discussed along with point estimation. We use a simulated data example and the two motivating examples to illustrate the proposed methodology. We also address certain issues such as standard error variability and comparison between multivariate and univariate analyses.

In Chapter 4 we consider inference in the multivariate GLMM. We describe hypothesis testing for the fixed effects based on approximations to the Wald and likelihood ratio statistics and propose score tests for the variance components and for testing the conditional independence assumption. The performance of the Gauss-Hermite quadrature and Monte Carlo approximations to the Wald, score and likelihood ratio statistics is compared via a small simulation study.

Chapter 5 introduces a correlated probit model as an alternative to the multivariate GLMM for binary and continuous data when conditional independence does not hold. We develop a Monte Carlo EM algorithm for maximum likelihood estimation and apply it to one of the motivating examples. To address the issue of advantages of joint over separate analyses of the response variables we design a simulation study

to investigate possible efficiency gains in a multivariate analysis. An identifiability issue concerning one of the variance components is also discussed.

The dissertation concludes with a summary of the most important findings and discussion of future research topics (Chapter 6).

CHAPTER 2
MULTIVARIATE GENERALIZED LINEAR MIXED MODEL

The Generalized Linear Mixed Model (GLMM) is a very general class of random effects models, well suited for subject-specific inference for repeated measures data. It is a special case of the random effects model defined in Section 1.2. We start this chapter by providing a detailed definition of the GLMM for repeated univariate mixed models (Section 2.1) and then introduce the multivariate extension (Section 2.2). Some properties of the multivariate GLMM are discussed in Section 2.3 and the chapter concludes with a description of two data sets which will be used for illustration throughout the thesis. Hereafter we omit the superscript c in the conditional mean μ_{ij}^c notation.

2.1 Introduction

Let y_{ij} denote the j^{th} response observed on the i^{th} subject, $j = 1, ...n_i$, $i = 1, ...n$. The conditional distribution of y_{ij} given an unobserved $q \times 1$ subject-specific random vector $\mathbf{b}_i$ is assumed to be in the exponential family with density

$$f(y_{ij}|\mathbf{b}_i) = exp\{\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\phi}\omega_{ij} + c(y_{ij}, \phi, \omega_{ij})\},$$

where $\mu_{ij} = b'(\theta_{ij})$ is the conditional mean, ϕ is the dispersion parameter, $b(\cdot)$ and $c(\cdot)$ are specific functions corresponding to the type of exponential family, and ω_{ij} are known weights. Also at this stage, it is assumed that $g(\mu_{ij}) = g(E(y_{ij}|\mathbf{b}_i)) = \eta_{ij}$, where $\eta_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i$ is a linear predictor, $\mathbf{b}_i$ is a $q \times 1$ random effect, $\boldsymbol{\beta}$ is a

$p \times 1$ parameter vector and the design vectors $\mathbf{x}_{ij}{}^{p\times 1}$ and $\mathbf{z}_{ij}{}^{q\times 1}$ are functions of the covariates.

At the second stage the subject-specific effects $\mathbf{b}_i$ are assumed to be i.i.d. $\mathbf{N_q}(\mathbf{0}, \boldsymbol{\Sigma})$. In general, the random effects can have other continuous or discrete distributions, but the normal distribution provides a full range of possible covariance structures and we will restrict our attention to this case.

As an additional assumption, conditional independence of the observations within and between subjects is required, that is,

$$f(\mathbf{y}|\mathbf{b};\boldsymbol{\beta},\phi) = \prod_{i=1}^{n} f(\mathbf{y}_i|\mathbf{b}_i;\boldsymbol{\beta},\phi) \quad \text{with} \quad \mathrm{f}(\mathbf{y}_i|\mathbf{b}_i;\boldsymbol{\beta},\phi) = \prod_{j=1}^{n_i} \mathrm{f}(y_{ij}|\mathbf{b}_i,\boldsymbol{\beta},\phi),$$

where $\mathbf{y} = (\mathbf{y}_1^T, ...\mathbf{y}_n^T)^T$ and $\mathbf{b} = (\mathbf{b}_1^T, ...\mathbf{b}_n^T)^T$ are column vectors of all responses and all random effects respectively.

Maximum-likelihood estimation for GLMM is complicated because the marginal likelihood for the response does not have a closed form expression. Different approaches for dealing with this situation are discussed in Chapter 3.

The parameters $\boldsymbol{\beta}$ in the GLMM have subject-specific interpretations; i.e. they describe the individual's rather than the average population response to changing the covariates. Under the GLMM, the marginal mean $\mu_{ij}^* = E(y_{ij}) = E(E(y_{ij}|\mathbf{b}_i))) = \int g^{-1}(\mathbf{x}_{ij}^T\boldsymbol{\beta}+\mathbf{z}_{ij}^T\mathbf{b}_i) f(\mathbf{b}_i, \boldsymbol{\Sigma}) d\mathbf{b}_i$ and in general $g(\mu_{ij}) \neq \mathbf{x}_{ij}^T\boldsymbol{\beta}$ if g is a non-linear function. However, this equation holds approximately if the standard deviations of the random effects distributions are small.

Closed-form expressions for the marginal means exist for certain link functions (Zeger, Liang and Albert, 1988). For example, for the identity link function the marginal and conditional moments coincide. For the log link function the marginal mean is $\mu_{ij}^* = exp(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \frac{\mathbf{z}_{ij}^T\boldsymbol{\Sigma}\mathbf{z}_{ij}}{2})$. For the probit link function, the marginal mean is $\mu_{ij}^* = \Phi(a_p(\boldsymbol{\Sigma})\mathbf{x}_{ij}^T\boldsymbol{\beta})$, where $a_p(\boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}\mathbf{z}_{ij}\mathbf{z}_{ij}^T + \mathbf{I}|^{-q/2}$ and q is the dimension of $\mathbf{b}_i$. For

the logit link, an exact closed-form expression for the marginal mean is unavailable but using a cumulative Gaussian approximation to the logistic function leads to the expression $logit(\mu_{ij}) \approx a_l(\boldsymbol{\Sigma})\mathbf{x}_{ij}^T\boldsymbol{\beta}$, where $a_l(\boldsymbol{\Sigma}) = |c^2\boldsymbol{\Sigma}\mathbf{z}_{ij}\mathbf{z}_{ij}^T + \mathbf{I}|^{-q/2}$ and $c = 16\frac{\sqrt{3}}{15\pi}$.

Unfortunately, there are no simple formulae for the higher order marginal moments except in the case of a linear link function. But the second-order moments can also be approximated and then the GEE approach can be used to fit the mixed model (Zeger, Liang and Albert, 1988).

2.2 Model Definition

Let us first consider bivariate repeated measures. Denote the response vector for the i^{th} subject by $\mathbf{y}_i = (\mathbf{y}_{i1}^T, \mathbf{y}_{i2}^T)^T$, where $\mathbf{y}_{i1} = (y_{i11}, ..., y_{i1n_i})^T$ and $\mathbf{y}_{i2} = (y_{i21}, ..., y_{i2n_i})^T$ are the repeated measurements on the 1^{st} and 2^{nd} variable respectively at n_i occasions. The number of observations for the two variables within a subject need not be the same and hence it would be more appropriate to denote them by n_{i1} and n_{i2} but for simplicity we will use n_i. We assume that y_{i1j}, $j = 1, ...n_i$, are conditionally independent given $\mathbf{b}_{i1}$ with density $f_1(\cdot)$ in the exponential family. Analogously, y_{i2j}, $j = 1, ...n_i$, are conditionally independent given $\mathbf{b}_{i2}$ with density $f_2(\cdot)$ in the exponential family. Note that f_1 and f_2 need not be the same. Also $\mathbf{y}_{i1}$ and $\mathbf{y}_{i2}$ are conditionally independent given $\mathbf{b}_i = (\mathbf{b}_{i1}^T, \mathbf{b}_{i2}^T)^T$ and the responses on different subjects are independent. Let $g_1(\cdot)$ and $g_2(\cdot)$ be appropriate link functions for f_1 and f_2. Denote the conditional means of y_{i1j} and y_{i2j} by μ_{i1j} and μ_{i2j} respectively. Let $\boldsymbol{\mu}_{i1} = (\mu_{i11}, ..., \mu_{i1n_i})^T$ and $\boldsymbol{\mu}_{i2} = (\mu_{i21}, ..., \mu_{i2n_i})^T$. At stage one of the mixed model specification we assume

$$g_1(\boldsymbol{\mu}_{i1}) = \mathbf{X}_{i1}\boldsymbol{\beta}_1 + \mathbf{Z}_{i1}\mathbf{b}_{i1} \tag{2.1}$$

$$g_2(\boldsymbol{\mu}_{i2}) = \mathbf{X}_{i2}\boldsymbol{\beta}_2 + \mathbf{Z}_{i2}\mathbf{b}_{i2}, \tag{2.2}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $p_1 \times 1$ and $p_2 \times 1$ dimensional unknown parameter vectors, $\mathbf{X}_{i1}$ and $\mathbf{X}_{i2}$ are $n_i \times p_1$ and $n_i \times p_2$ dimensional design matrices for the fixed effects, $\mathbf{Z}_{i1}$ and $\mathbf{Z}_{i2}$ are $n_i \times q_1$ and $n_i \times q_2$ design matrices for the random effects and g_1 and g_2 are applied componentwise to $\boldsymbol{\mu}_{i1}$ and $\boldsymbol{\mu}_{i2}$. At stage two, a joint distribution for $\mathbf{b}_{i1}$ $(q_1 \times 1)$ and $\mathbf{b}_{i2}$ $(q_2 \times 1)$ is specified. The normal distribution is a very good candidate, as it provides a rich covariance structure. Hence, we assume

$$\mathbf{b}_i = \begin{pmatrix} \mathbf{b}_{i1} \\ \mathbf{b}_{i2} \end{pmatrix} \sim i.i.d.\, \mathbf{MVN}(\mathbf{0}, \boldsymbol{\Sigma}) = \mathbf{MVN}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right), \tag{2.3}$$

where $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ are in general unknown positive-definite matrices. The extension of this model to higher dimensional multivariate repeated measures is straightforward but for simplicity of discussion only the bivariate case will be considered.

When $\boldsymbol{\Sigma}_{12} = \mathbf{0}$ then the above model is equivalent to two separate GLMM's for the two outcome variables. Advantages of joint over separate fitting include the ability to answer intrinsically multivariate questions, better control over the type I error rates in multiple tests and possible gains in efficiency in the parameter estimates.

For example, in the developmental toxicity application described in more detail at the end of this section, it is of interest to estimate the dose effect of ethylene glycol on both outcomes (malformation and low fetal weight) simultaneously. One might be interested in the probability of either malformation or fetal weight at any given dose. This type of question can not be answered by a univariate analysis as it requires knowledge of the correlation between the outcomes. Also, testing the significance of the dose effect on malformation and fetal weight simultaneously (using a Wald test for example) allows one to keep the significance level fixed at α. While if the two outcomes were tested separately then an adjustment should be made for the level of the individual tests to achieve overall α level for both comparisons.

Multivariate analysis also allows one to borrow strength from the observations on one outcome to estimate the other outcomes. This can lead to more precise estimates

of the parameters and therefore, to gains in efficiency. Theoretically, the gain will be the greatest if there is a common random effect for all responses (this means $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{22}$ in our notation). Also, if the number of observations per subject is small and the correlation between the two outcomes is large, the gains in efficiency should increase.

If the two vectors of random effects are perfectly correlated, then this is equivalent to assuming a common latent variable with a multivariate normal distribution which does not depend on covariates. Hence, the multivariate GLMM reduces to a special case of the Sammel, Ryan and Legler model when the latent variable is not modelled as a function of the covariates. It is possible to allow the random effects in our models to depend on covariates but we have not considered that extension here.

Several other models discussed in previous sections turn out to be special cases of this general formulation. For example the bivariate Rasch model is obtained by specifying Bernoulli response distributions for both variables, using the logit link function and identity design matrices both for the fixed and for the random effects. The multivariate binomial-logit normal model is also a special case for binary responses with identity design matrix for the random effects and unrestricted variance-covariance structure. The Aitchison and Ho multivariate Poisson log-normal model also falls under this general structure when the response variables are assumed to have a Poisson distribution.

2.3 Model Properties

Exactly as in GLMM, the conditional moments in the multivariate GLMM are directly modelled, while marginal moments are harder to find. The marginal means and the marginal variances of $\mathbf{y}_{i1}$ and $\mathbf{y}_{i2}$ for the model defined by (2.2) and (2.3) are the same as those of the GLMM considering one variable at a time:

$$E(\mathbf{y}_{i1}) = EE(\mathbf{y}_{i1}|\mathbf{b}_{i1}) = E[\boldsymbol{\mu}_{i1}(\boldsymbol{\beta}_1, \mathbf{b}_{i1})],$$

$$E(\mathbf{y}_{i2}) = E[\boldsymbol{\mu}_{i2}(\boldsymbol{\beta}_2, \mathbf{b}_{i2})],$$

$$Var(\mathbf{y}_{i1}) = E[Var(\mathbf{y}_{i1}|\mathbf{b}_{i1})] + Var[E(\mathbf{y}_{i1}|\mathbf{b}_{i1})] =$$

$$E[\phi_1 V(\boldsymbol{\mu}_{i1})] + Var[\boldsymbol{\mu}_{i1}],$$

$$Var(\mathbf{y}_{i2}) = E[\phi_2 V(\boldsymbol{\mu}_{i2})] + Var[\boldsymbol{\mu}_{i2}],$$

where $V(\boldsymbol{\mu}_{i1})$ and $V(\boldsymbol{\mu}_{i2})$ denote the variance functions corresponding to the exponential family distributions for the two response variables.

The marginal covariance matrix between $\mathbf{y}_{i1}$ and $\mathbf{y}_{i2}$ is found to be equal to the covariance matrix between the conditional means $\boldsymbol{\mu}_{i1}$ and $\boldsymbol{\mu}_{i2}$:

$$Cov(\mathbf{y}_{i1}, \mathbf{y}_{i2}) = E(\mathbf{y}_{i1}\mathbf{y}_{i2}^T) - E(\mathbf{y}_{i1})E(\mathbf{y}_{i2}^T) =$$

$$= EE(\mathbf{y}_{i1}\mathbf{y}_{i2}^T|\mathbf{b}_{i1}, \mathbf{b}_{i2}) - E(\boldsymbol{\mu}_{i1})E(\boldsymbol{\mu}_{i2}^T) =$$

$$= E[E(\mathbf{y}_{i1}|\mathbf{b}_{i1})E(\mathbf{y}_{i2}^T|\mathbf{b}_{i2})] - E(\boldsymbol{\mu}_{i1})E(\boldsymbol{\mu}_{i2}^T) =$$

$$= Cov(\boldsymbol{\mu}_{i1}, \boldsymbol{\mu}_{i2})$$

The latter property is a consequence of the key assumption of conditional independence between the two response variables. This assumption allows one to extend model fitting methods from the univariate to the multivariate GLMM but may not hold in certain situations. This issue will be discussed in more detail in Chapter 4, where score tests are proposed for verifying conditional independence.

2.4 Applications

The multivariate GLMM are fitted to two data sets. The first one is a dataset from a developmental toxicity study of ethylene glycol (EG) in mice conducted through the National Toxicology Program (Price et al., 1985). The experiment involved four randomly chosen groups of pregnant mice, one group serving as a control and the other three exposed to three different levels of EG during major organogenesis. Following

Table 2.1. Descriptive statistics for the Ethylene Glycol data

Dose (g/kg)	Dams	Live Fetuses	Fetal Weight (g)		Malformation	
			Mean	SD	Number	Percent
0	25	297	0.972	0.098	1	0.34
0.75	24	276	0.877	0.104	26	9.42
1.50	22	229	0.764	0.107	89	38.86
3.00	23	226	0.704	0.124	126	57.08

sacrifice, measurements were taken on each fetus in the uterus. The two outcome measures on each live fetus of interest to us are fetal weight (continuous) and malformation status (dichotomous). Some descriptive statistics for the data are available in Table 2.1. Fetal weight decreases monotonically with increasing dose with the average weight ranging from 0.972 g in the control group to 0.704 g in the group administered the highest dose. At the same time the malformation rate increases with dose from 0.3% in the control group to 57% in the group administered the highest dose.

The goal of the analysis is to study the joint effects of increasing dose on fetal weight and on the probability of malformation. The analysis of these data is complicated by the correlations between the repeated measures on fetuses within litter. A multivariate GLMM with random intercepts for each variable allows one to explicitly model the correlation structure within litter and provides subject-specific estimates for the regression parameters.

The second data set is from a study to compare the effects of 9 drugs and a placebo on the patterns of myoelectric activity in the intestines of ponies (Lester et al., 1998a, 1998b, 1998c). For that purpose electrodes are attached to four different areas of the intestines of 6 ponies, and spike burst rate and duration are measured at 18 equally spaced time intervals around the time of each drug administration. Six of the drugs and the placebo are given twice to each pony in a randomized complete block design. The remaining three drugs are not given to all ponies and hence will not be analyzed here. There is a rest period after each drug's administration and no carry-over effects

Table 2.2. Descriptive statistics for pony data

	Duration		Count		
Pony	Mean	SD	Mean	SD	Corr.
1	1.03	0.25	77.57	45.53	0.59
2	1.35	0.28	128.25	76.08	0.18
3	1.40	0.36	84.33	46.27	0.45
4	1.27	0.48	111.00	49.66	0.21
5	1.14	0.21	75.44	45.03	0.50
6	1.24	0.40	66.86	48.97	0.32

are expected. The spike burst rate is a count variable reflecting the number of contractions exceeding certain threshold in 15 minutes. The duration variable reflects average duration of the contractions in each 15-minute interval. Figure 2.1 shows graphical representations of the averaged responses by drug and time for one of the electrodes and one hour after the drug administration. Table 2.2 shows the sample means, standard deviations and correlations between the two outcome variables by pony for this smaller data set. We analyse a restricted data set for reasons of computational and practical feasibility. This issue is discussed in more detail in Chapter 3.

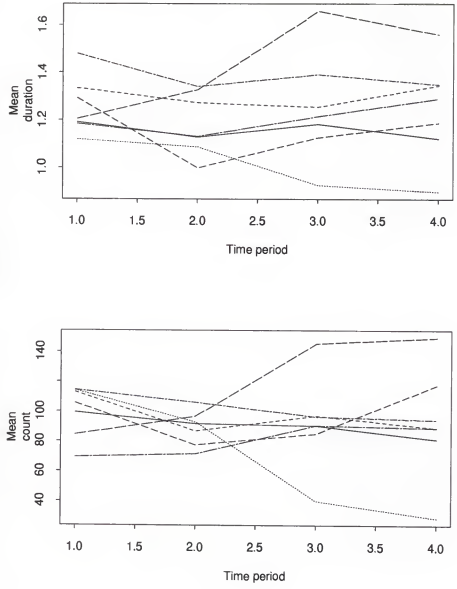

Figure 2.1 Count and duration trends over time for the pony data. Each trajectory shows the change in mean response for one of the seven drugs.

CHAPTER 3
ESTIMATION IN THE MULTIVARIATE GENERALIZED LINEAR MIXED MODEL

This chapter focuses on methods for obtaining maximum likelihood (or approximate maximum likelihood) estimates for the model parameters in the multivariate GLMM. We first mention some approaches proposed for the univariate GLMM (Section 3.1), and then describe in detail extensions of three of those approaches for the multivariate GLMM (Section 3.2). The proposed methods are then illustrated on a simulated data example (Section 3.3), and on the two 'real-life' data sets introduced in Chapter 2 (Section 3.4). Some issues such as standard error variability and advantages of one multivariate versus several univariate analyses are addressed in Sections 3.3 and 3.4. The chapter concludes with discussion of some additional model-fitting methods (Section 3.5).

3.1 Introduction

In GLMM the marginal likelihood of the response is obtained by integrating out the random effects

$$\prod_{i=1}^{n} \int f(\mathbf{y}_i|\mathbf{b}_i;\boldsymbol{\beta},\phi) f(\mathbf{b}_i) d\mathbf{b}_i,$$

where $f(\mathbf{b}_i)$ denotes a normal density. Usually these integrals are not analytically tractable and some kind of approximation must be used.

Direct approaches to maximum likelihood estimation are based on numerical or stochastic approximations of the integrals and on numerical maximizations of those approximations. Probably the most widely used numerical approximation procedure

for this type of integral is Gauss-Hermite quadrature. It involves evaluating the integrands at m prespecified quadrature points and substituting weighted sums in place of the intractable integrals for the n subjects. If the number of quadrature points m is large the approximation can be made very accurate but to keep the numerical effort low m should be kept as small as possible. Gauss-Hermite quadrature is appropriate when the dimension of the random effects is small. An alternative for high-dimensional random effects is to approximate the integrals by Monte Carlo sums. This involves generating m random values from the random effects distribution for each subject, evaluating the conditional densities $f(\mathbf{y}_i|\mathbf{b}_i;\boldsymbol{\beta},\phi)$ at those values and taking averages. Details on how to perform Gauss-Hermite quadrature and Monte Carlo approximation for the GLMM can be found in Fahrmeir and Tutz (1994, pp.357-365). We discuss Gauss-Hermite quadrature for the multivariate GLMM in Section 3.2. Liu and Pierce (1994) consider adaptive Gaussian quadrature, which allows a reduction in the required number of quadrature points by centering and scaling them around the mode of the integrand function. This procedure is described in more detail and applied to one of the data examples in Section 3.4.

Indirect approaches to maximum likelihood estimation use the EM-algorithm (Dempster, Laird and Rubin, 1977), treating the random effects as the missing data. We apply these methods both to the multivariate GLMM and to the correlated probit model and therefore we now introduce the basic ideas. Hereafter $\boldsymbol{\psi}$ denotes the vector of all unknown parameters in the problem, and $\hat{\boldsymbol{\psi}}$ denotes some estimate of $\boldsymbol{\psi}$. The EM algorithm is an iterative technique for finding maximum likelihood estimates when direct maximization of the observed likelihood $f(\mathbf{y};\boldsymbol{\psi})$ is not feasible. It involves augmenting the observed data by unobserved data so that maximization at each step of the algorithm is considerably simplified. The unobserved data are denoted by $\mathbf{b}$ and in the GLMM context these are the random effects. The EM-algorithm can be summarized as follows:

1. Select a starting value $\hat{\psi}^{(0)}$. Set $r = 0$.

2. Increase r by 1.

E-step: Calculate $E\{ln f(\mathbf{y}, \mathbf{b}; \psi)|\mathbf{y}; \hat{\psi}^{(r-1)}\}$.

3. **M-step:** Find a value $\hat{\psi}^{(r)}$ of ψ that maximizes this conditional expectation.

4. Iterate between (2) and (3) until convergence is achieved.

In the GLMM context the complete data is $\mathbf{u} = (\mathbf{y}^T, \mathbf{b}^T)^T$ and the complete data log-likelihood is given by

$$ln\, \mathbf{L}_u = \sum_{i=1}^{n} ln\, f(\mathbf{y}_i|\mathbf{b}_i; \boldsymbol{\beta}, \phi) + \sum_{i=1}^{n} ln\, f(\mathbf{b}_i|\boldsymbol{\Sigma}).$$

The r^{th} E-step of the EM algorithm involves computing $E(ln\mathbf{L}_u|\mathbf{y}, \hat{\psi}^{(r-1)})$ and the r^{th} M-step maximizes this quantity with respect to ψ and updates the parameter estimates. Notice that because $\boldsymbol{\beta}$ and ϕ enter only the first term of $\mathbf{L}_u$, the M-step with respect to $\boldsymbol{\beta}$ and ϕ uses only $f(\mathbf{y}|\mathbf{b})$ and so it is similar to a standard generalized linear model computation with the values of $\mathbf{b}$ treated as known. Maximizing with respect to $\boldsymbol{\Sigma}$ is just maximum likelihood using the distribution of $\mathbf{b}$ after replacing sufficient statistics with their conditional expected values. In general, the conditional expectations in the E-step can not be computed in closed form, but Gauss-Hermite or different Monte Carlo approximations can be utilized. (Fahrmeir and Tutz, 1994, pp. 362-365; McCulloch, 1997; Booth and Hobert, 1999).

Many authors consider tractable analytical approximations to the likelihood (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1994). Although these methods lead to inconsistent estimates in some cases, they may have considerable advantage in computational speed over 'exact' methods and can be fitted with standard software. Both Breslow and Clayton's and Wolfinger and O'Connell's procedures amount to iterative fitting of normal theory linear mixed models and can be implemented using the

%*GLIMMIX* macro in SAS. An extension of Wolfinger and O'Connell's approach is considered in Section 3.2.

A Bayesian paradigm with flat priors can also be used to approximate the maximum likelihood estimates using posterior modes (Fahrmeir and Tutz, 1994, pp.233-238) or posterior means (Zeger and Karim, 1991). Though the numerator in such computations is the same as for the maximum likelihood calculations, the posterior may not exist for diffuse priors (Natarajan and McCulloch, 1995). This may not be detected using computational techniques such as the Gibbs sampler, and can result in incorrect parameter estimates (Hobert and Casella, 1996).

3.2 Maximum Likelihood Estimation

For simplicity of presentation we again consider the case of only two response variables. The marginal likelihood in the bivariate GLMM is obtained as in the usual GLMM by integrating out the random effects

$$\prod_{i=1}^{n} \int \int \{ \prod_{j=1}^{n_i} f_1(y_{i1j}|\mathbf{b}_{i1}; \boldsymbol{\beta}_1, \phi_1) f_2(y_{i2j}|\mathbf{b}_{i2}; \boldsymbol{\beta}_2, \phi_2) \} f(\mathbf{b}_{i1}, \mathbf{b}_{i2}; \boldsymbol{\Sigma}) d\mathbf{b}_{i1} d\mathbf{b}_{i2}, \qquad (3.1)$$

where f denotes the multivariate normal density of the random effects. In this section we describe how Gauss-Hermite quadrature, Monte Carlo EM algorithm and pseudo-likelihood can be used to obtain estimates in the multivariate GLMM. Both Gaussian quadrature and the Monte Carlo EM algorithm are referred to as 'exact maximum likelihood' methods because they target the exact maximum likelihood estimates. In comparison, methods that use analytical approximations are referred to as 'approximate' maximum likelihood. We now start by describing the 'exact' maximum likelihood methods. Hereafter $\boldsymbol{\psi} = Vec(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \phi_1, \phi_2, \boldsymbol{\delta})$, where $\boldsymbol{\delta}$ are unknown variance components in $\boldsymbol{\Sigma}$.

3.2.1 Gauss-Hermite Quadrature

Parameter Estimation

The marginal log-likelihood in the multivariate GLMM is expressed as a sum of the individual log-likelihoods for all subjects, that is

$$lnL(\mathbf{y};\boldsymbol{\psi}) = \sum_{i=1}^{n} lnL_i(\boldsymbol{\psi}),$$

where

$$L_i(\boldsymbol{\psi}) = \int_{\mathbf{R}^q} \{\prod_{j=1}^{n_i} f_1(y_{i1j}|\mathbf{b}_{i1};\boldsymbol{\beta}_1,\phi_1) f_2(y_{i2j}|\mathbf{b}_{i2};\boldsymbol{\beta}_2,\phi_2)\} \frac{1}{|2\pi\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp(-\frac{1}{2}\mathbf{b}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{b}_i) d\mathbf{b}_i.$$

Hence Gauss-Hermite quadrature (Fahrmeir and Tutz, 1994) involves numerical approximations of n q-dimensional integrals ($q = q_1 + q_2$). The $\mathbf{b}_i$'s are first transformed so that each integral has the form

$$L_i = \frac{1}{\pi} \int_{\mathbf{R}^q} h(\mathbf{z}_i) exp(-\mathbf{z}_i^T\mathbf{z}_i) d\mathbf{z}_i.$$

The needed transformation is $\mathbf{b}_i = \sqrt{2}\mathbf{L}\mathbf{z}_i$, where $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$. $\mathbf{L}$ is the lower triangular Choleski factor of $\boldsymbol{\Sigma}$ and it always exists because $\boldsymbol{\Sigma}$ is non-negative definite. Here

$$h(\mathbf{z}_i) = \prod_{j=1}^{n_i} f_1(y_{i1j}|\mathbf{z}_{i1};\boldsymbol{\beta}_1,\phi_1) f_2(y_{i2j}|\mathbf{z}_{i2};\boldsymbol{\beta}_2,\phi_2).$$

Each integral is then approximated by

$$L_i^{GQ} = \sum_{k_1=1}^{m} \nu_{k_1}^{(1)} ... \sum_{k_q=1}^{m} \nu_{k_q}^{(q)} h(\mathbf{z}^{(k)}),$$

where $\mathbf{z}^{(k)} = \sqrt{2}\mathbf{L}\mathbf{d}^{(k)}$ for the multiple index $\mathbf{k} = (k_1, ...k_q)$ and $\mathbf{d}^{(k)}$ denote the tabled nodes of univariate Gauss-Hermite integration of order m (Abramowitz and Stegun,

1972). The corresponding weights are given by $\nu_{k_l}^{(l)} = \pi^{-\frac{1}{2}} w_{k_l}^{(l)}$, where $w_{k_l}^{(l)}$ are the tabled univariate weights, $l = 1, ...m$.

The maximization algorithm then proceed as follows:

1. Choose initial estimate for the parameter vector $\hat{\psi}^{(0)}$. Set $r = 0$

2. Increase r by 1.

Approximate each of the integrals $L_i(\hat{\psi}^{(r-1)})$ by $L_i^{GQ}(\hat{\psi}^{(r-1)})$ using m quadrature points in each direction.

3. Maximize the approximation with respect to ψ using a numerical maximization routine.

4. Iterate between steps (2) and (3) until the parameter estimates have converged.

A popular numerical maximization procedure for step 3 is the Newton-Raphson method. It involves iteratively solving the equations

$$\hat{\psi}^{(r)} = \hat{\psi}^{(r-1)} + [\mathbf{J}(\hat{\psi}^{(r-1)})]^{-1}\mathbf{S}(\hat{\psi}^{(r-1)}),$$

where $\mathbf{S}(\hat{\psi}^{(r)}) = \frac{\partial \ln L}{\partial \psi}|_{\hat{\psi}^{(r-1)}}$, and $\mathbf{J}(\hat{\psi}^{(r-1)}) = -\frac{\partial^2 \ln L}{\partial \psi \partial \psi'}|_{\hat{\psi}^{(r-1)}}$ is the observed information matrix. One possible criterion for convergence is

$$\max_s \frac{|\hat{\psi}_s^{(r+1)} - \hat{\psi}_s^{(r)}|}{|\hat{\psi}_s^{(r)}| + \delta_2} < \delta_1,$$

where $\hat{\psi}_s^{(r)}$ denotes the estimate of the s^{th} element of the parameter vector at step r of the algorithm and δ_1 and δ_2 are chosen to be small positive numbers. The role of δ_2 is to prevent numerical problems stemming from estimates close to zero. Another frequently used criterion is

$$\| \hat{\psi}^{(r+1)} - \hat{\psi}^{(r)} \| < \delta,$$

where $\| \cdot \|$ denotes Euclidean norm.

The numerical maximization procedure MaxBFGS which we are going to use for the numerical examples later in this chapter is based on two criteria for convergence:

$$|s_s^{(r)}\hat{\psi}_s^{(r)}| \leq \epsilon \text{ for all } s \text{ when } \hat{\psi}_s^{(r)} \neq 0$$

$$|s_s^{(r)}| \leq \epsilon \text{ for all } s \text{ when } \hat{\psi}_s^{(r)} = 0$$

and

$$|\hat{\psi}_s^{(r+1)} - \hat{\psi}_s^{(r)}| \leq 10\epsilon|\hat{\psi}_s^{(r)}| \text{ for all } s \text{ when } \hat{\psi}_s^{(r)} \neq 0$$

$$|\hat{\psi}_s^{(r+1)} - \hat{\psi}_s^{(r)}| \leq 10\epsilon \text{ for all } s \text{ when } \hat{\psi}_s^{(r)} = 0$$

where s_s denotes the s^{th} component of the score vector.

Standard Error Estimation

After the algorithm has converged estimates of the standard errors of the parameter estimates can be based on the observed information matrix. By asymptotic maximum-likelihood theory

$$Var(\hat{\boldsymbol{\psi}}) = \mathbf{I}^{-1}(\boldsymbol{\psi}),$$

where $\mathbf{I}(\boldsymbol{\psi}) = E(-\frac{\partial^2 lnL}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'})$ is the expected information matrix. But the observed information matrix $\mathbf{J}(\boldsymbol{\psi})$ is easier to obtain and hence we will use the latter to approximate the standard errors

$$s.e.(\hat{\psi}_s) = \sqrt{i^{ss}},$$

where i^{ss} is the s^{th} diagonal element of the inverse of the observed information matrix. Note that if the Newton-Raphson maximization method is used the observed information matrix is a by-product of the algorithm.

Numerical and Exact Derivatives

The observed information matrix and the score vector are not available in closed-form and must be approximated. One can either compute numerical derivatives of the approximated log-likelihood, or approximate the intractable integrals in the expressions for the exact derivatives. The first approach is simpler to implement but may require a large number of quadrature points.

The method of finite differences can be used to find numerical derivatives. The s^{th} element of the score vector and the $(s,t)^{th}$ element of the information matrix are approximated as follows:

$$s_s(\boldsymbol{\psi}) \approx \sum_{i=1}^{n} \frac{l_i^a(\boldsymbol{\psi} + \epsilon \imath_s) - l_i^a(\boldsymbol{\psi} - \epsilon \imath_s)}{2\epsilon}$$

$$i_{st}(\boldsymbol{\psi}) \approx$$

$$\sum_{i=1}^{n} \frac{l_i^a(\boldsymbol{\psi} + \epsilon_1 \imath_s + \epsilon_2 \imath_t) - l_i^a(\boldsymbol{\psi} - \epsilon_1 \imath_s + \epsilon_2 \imath_t) - l_i^a(\boldsymbol{\psi} + \epsilon_1 \imath_s - \epsilon_2 \imath_t) + l_i^a(\boldsymbol{\psi} - \epsilon_1 \imath_s - \epsilon_2 \imath_t)}{4\epsilon_1 \epsilon_2},$$

where $l_i^a = lnL_i^{GQ}$, ϵ_1 and ϵ_2 are suitably chosen step lengths, and $\imath_s$ and $\imath_t$ are unit vectors. The unit vectors have ones in positions s and t respectively and the rest of their elements are zeros.

An alternative to numerical derivatives is to approximate the exact derivatives. The latter are obtained by interchanging the differentiation and integration signs and are as follows:

$$\frac{\partial ln\ L}{\partial \boldsymbol{\psi}} = \sum_{i=1}^{n} \int \frac{\partial ln f_i}{\partial \boldsymbol{\psi}} f_i(\mathbf{y}_i, \mathbf{b}_i | \boldsymbol{\psi}) d\mathbf{b}_i$$

and

$$\frac{\partial^2 ln\ L}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T} = \sum_{i=1}^{n} \frac{\int (\frac{\partial^2 ln f_i}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T} + \frac{\partial ln f_i}{\partial \boldsymbol{\psi}} \frac{\partial ln f_i}{\partial \boldsymbol{\psi}^T}) f_i d\mathbf{b}_i \int f_i d\mathbf{b}_i - \int \frac{\partial ln f_i}{\partial \boldsymbol{\psi}} f_i d\mathbf{b}_i \int \frac{\partial ln f_i}{\partial \boldsymbol{\psi}^T} f_i d\mathbf{b}_i}{(\int f_i d\mathbf{b}_i)^2},$$

where

$$f_i = \prod_{j=1}^{n_i} \{f_1(y_{i1j}|\mathbf{b}_i;\boldsymbol{\beta}_1,\phi_1) f_2(y_{i2j}|\mathbf{b}_i;\boldsymbol{\beta}_2,\phi_2)\} f(\mathbf{b}_i;\boldsymbol{\Sigma}).$$

For each subject there are three intractable integrals which need to be approximated: $\int f_i d\mathbf{b}_i$, $\int \frac{\partial ln f_i}{\partial \psi} f_i d\mathbf{b}_i$ and $\int (\frac{\partial^2 ln f_i}{\partial \psi \partial \psi^T} + \frac{\partial ln f_i}{\partial \psi} \frac{\partial ln f_i}{\partial \psi}) f_i d\mathbf{b}_i$, and Gauss-Hermite quadrature is used for each one of them. Note that

$$ln f_i = \sum_{i=1}^{n} ln f_1(y_{i1j}|\mathbf{b}_i;\boldsymbol{\beta}_1,\phi_1) + \sum_{i=1}^{n} ln f_2(y_{i2j}|\mathbf{b}_i;\boldsymbol{\beta}_2,\phi_2) + ln f(\mathbf{b}_i;\boldsymbol{\Sigma}).$$

Hence $\frac{\partial ln f_i}{\partial \boldsymbol{\beta}_1} = \frac{\partial (\sum_{j=1}^{n_i} ln f_1)}{\partial \boldsymbol{\beta}_1}$, $\frac{\partial ln f_i}{\partial \boldsymbol{\beta}_2} = \frac{\partial (\sum_{j=1}^{n_i} ln f_2)}{\partial \boldsymbol{\beta}_2}$ and $\frac{\partial ln f_i}{\partial \sigma_{st}} = \frac{\partial ln f}{\partial \sigma_{st}}$. Similarly the derivatives with respect to the scale parameters ϕ_1 and ϕ_2, if needed, are obtained by differentiating the first and second term in the summation for $ln f_i$. This leads to simple expressions for the integrands because the functions that are differentiated are members of the exponential family (Fahrmeir and Tutz, 1994).

Note that the random effects distribution is always multivariate normal which complicates the expressions for the second order derivatives with respect to the variance components. Jennrich and Schluster (1986) propose an elegant approach to deal with this problem and we now illustrate their method for the multivariate GLMM. For simplicity of notation consider bivariate random effects $\mathbf{b}_i$. Let $\boldsymbol{\delta} = \begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{pmatrix} = \begin{pmatrix} \sigma_1 \\ \sigma_2 \\ \rho \end{pmatrix}$ and let $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. Then write

$$\boldsymbol{\Sigma} = \sigma_1^2 \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \sigma_2^2 \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} + \rho\sigma_1\sigma_2 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Let $\dot{\boldsymbol{\Sigma}}_s = \frac{\partial \boldsymbol{\Sigma}}{\partial \delta_s}$, $s = 1, ...3$. Then

$$\frac{\partial ln f_i}{\partial \delta_s} = \frac{1}{2} tr\{\boldsymbol{\Sigma}^{-1}(\mathbf{b}_i\mathbf{b}_i^T - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_s\}$$

and

$$\frac{\partial^2 ln f_i}{\partial \delta_s \partial \delta_t} = -\frac{1}{2} tr\{\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_s\boldsymbol{\Sigma}^{-1}\dot{\boldsymbol{\Sigma}}_t\}.$$

In the above parametrization of the variance-covariance matrix there are restrictions on the parameter space ($\sigma_1 > 0$, $\sigma_2 > 0$, $-1 < \rho < 1$), but the algorithm does not guarantee that the current estimates will stay in the restricted parameter space. It is then preferable to maximize with respect to the elements of the Choleski root of $\boldsymbol{\Sigma}$. The Choleski root is a left-triangular matrix $\mathbf{L} = \begin{pmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{pmatrix}$ such that $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$. Its elements are held unrestricted in intermediate steps of the algorithm and hence they can lead to a negative-definite final estimate of $\boldsymbol{\Sigma}$. Such an outcome indicates either a numerical problem in the procedure or an unappropriate model. However, if only intermediate estimates of $\boldsymbol{\Sigma}$ are negative-definite, then the maximization algorithm with respect to the elements of the Choleski root will work well while the other one can have problems.

Although the parametrization in terms of l_{11}, l_{21} and l_{22} has fewer numerical problems than the parametrization in terms of σ_1, σ_2 and ρ, it does not have nice interpretation. It then seems reasonable to use the Choleski roots in the maximization procedure but to perform one extra partial step of the algorithm to approximate the information matrix for the meaningful parametrization σ_1, σ_2 and ρ.

3.2.2 Monte Carlo EM Algorithm

Parameter Estimation

The complete data for the multivariate GLMM is $\mathbf{u} = (\mathbf{y}^T, \mathbf{b}^T)^T$ and hence the complete data log-likelihood is given by

$$ln\mathbf{L}_u = \sum_{i=1}^{n}\sum_{j_1=1}^{n_i} ln\, f_1(y_{i1j_1}|\mathbf{b}_{i1};\boldsymbol{\beta}_1,\phi_1) + \sum_{i=1}^{n}\sum_{j_2=1}^{n_i} ln\, f_2(y_{i2j_2}|\mathbf{b}_{i2};\boldsymbol{\beta}_2,\phi_2) + \sum_{i=1}^{n} ln\, f(\mathbf{b}_i,\boldsymbol{\Sigma}).$$

Notice that $(\boldsymbol{\beta}_1, \phi_1)$, $(\boldsymbol{\beta}_2, \phi_2)$ and $\boldsymbol{\Sigma}$ appear in different terms in the log-likelihood and therefore each M-step of the algorithm consists of three separate maximizations of $E[\sum_{i=1}^{n}\sum_{j_1=1}^{n_i} ln\, f_1(y_{i1j_1}|\mathbf{b}_{i1};\boldsymbol{\beta}_1,\phi_1)|\mathbf{y}]$, $E[\sum_{i=1}^{n}\sum_{j_2=1}^{n_i} ln\, f_2(y_{i2j_1}|\mathbf{b}_{i2};\boldsymbol{\beta}_2,\phi_2)|\mathbf{y}]$, and $E[\sum_{i=1}^{n} ln\, f(\mathbf{b}_i,\boldsymbol{\Sigma})|\mathbf{y}]$ respectively, evaluated at the current parameter estimate of $\boldsymbol{\psi}$. Recall that for the regular GLMM there are two separate terms to be maximized.

These conditional expectations do not have closed form expressions but can be approximated by Monte Carlo estimates:

$$\frac{1}{m}\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j_1=1}^{n_i} ln\, f_1(y_{i1j_1}|\mathbf{b}_{i1}^{(k)};\boldsymbol{\beta}_1,\phi_1) \tag{3.2}$$

$$\frac{1}{m}\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{j_2=1}^{n_i} ln\, f_2(y_{i2j_2}|\mathbf{b}_{i2}^{(k)};\boldsymbol{\beta}_2,\phi_2) \tag{3.3}$$

$$\frac{1}{m}\sum_{k=1}^{m}\sum_{i=1}^{n} ln\, f(\mathbf{b}_i^{(k)};\boldsymbol{\Sigma}), \tag{3.4}$$

where $(\mathbf{b}_{i1}^{(k)}, \mathbf{b}_{i2}^{(k)})$, $k = 1, ...m$ are simulated values from the distribution of $\mathbf{b}_i|\mathbf{y}_i$, evaluated at the current parameter estimate of $\boldsymbol{\psi}$. In order to obtain simulated values for $\mathbf{b}_i$ multivariate rejection sampling or importance sampling can be used as proposed by Booth and Hobert (1999). At the r^{th} step of the Monte Carlo EM algorithm multivariate rejection sampling algorithm is as follows:

For each subject i, $i = 1, ...n$,

1. Compute $\tau_i = \sup_{\mathbf{b}_i \in \Re^q} f(\mathbf{y}_i|\mathbf{b}_i, \hat{\boldsymbol{\psi}}^{(r-1)})$.

2. Sample $\mathbf{b}_i^{(k)}$ from the multivariate normal density $f(\mathbf{b}_i; \hat{\boldsymbol{\Sigma}}^{(r-1)})$ and independently sample $w_i^{(k)}$ from $Uniform(0,1)$.

3. If $w_i^{(k)} \leq \frac{f(\mathbf{y}_i|\mathbf{b}_i^{(k)})}{\tau_i}$, then accept $\mathbf{b}_i^{(k)}$; if not, go to 2.

Iterate between (2) and (3) until m generated values of $\mathbf{b}_i$ are accepted.

Once m simulated values for the random effects are available, the Monte Carlo estimates of the conditional expectations can be maximized using some numerical

maximization procedure such as the Newton-Raphson algorithm. Convergence can be claimed when successive values for the parameters are within some required precision. The same convergence criterions as those mentioned for Gauss-Hermite quadrature can be applied but with generally larger δ_1 value. Booth and Hobert (1999) suggest using δ_1 between 0.002 and 0.005. This is needed to avoid excessively large simulation sample sizes.

A nice feature of the Booth and Hobert algorithm is that because independent random sampling is used, Monte Carlo error can be assessed at each iteration and the simulation sample size can be automatically increased. The idea is that it is inefficient to start with a large simulation sample but as the estimates get closer to the maximum likelihood estimates the simulation sample size must be increased to provide enough precision for convergence.

Booth and Hobert propose to increase the simulation sample size m with m/t, where t some positive number, if $\hat{\boldsymbol{\psi}}^{(r-1)}$ lies in a 95% confidence ellipsoid constructed around $\hat{\boldsymbol{\psi}}^{(r)}$,that is if

$$(\hat{\boldsymbol{\psi}}^{(r)} - \hat{\boldsymbol{\psi}}^{(r-1)})^T \{\hat{var}(\hat{\boldsymbol{\psi}}^{(r)}|\hat{\boldsymbol{\psi}}^{(r-1)})\}^{-1} (\hat{\boldsymbol{\psi}}^{(r)} - \hat{\boldsymbol{\psi}}^{(r-1)}) < c^2,$$

where c^2 denotes the 95^{th} percentile of Chi-square distribution with number of degrees of freedom equal to the dimension of the parameter vector. To describe what variance approximation is used above some notation must be introduced. Denote the maximizer of the true conditional expectation $Q = E(ln\, \mathbf{L}_u(\mathbf{b})|\mathbf{y}, \hat{\boldsymbol{\psi}}^{(r-1)})$ by $\hat{\boldsymbol{\psi}}^{*(r)}$. Then

$$Var(\hat{\boldsymbol{\psi}}^{(r)}|\hat{\boldsymbol{\psi}}^{(r-1)}) \approx Q_m^{(2)}(\hat{\boldsymbol{\psi}}^{*(r)}|\hat{\boldsymbol{\psi}}^{(r-1)})^{-1} var\{(Q_m^{(1)}(\hat{\boldsymbol{\psi}}^{*(r)}|\hat{\boldsymbol{\psi}}^{(r-1)})\} Q_m^{(2)}(\hat{\boldsymbol{\psi}}^{*(r)}|\hat{\boldsymbol{\psi}}^{(r-1)})^{-1}.$$

Here $Q_m = \frac{1}{m}\sum_{k=1}^{m} ln\, \mathbf{L_u}(\mathbf{b}^{(k)}|\mathbf{y}, \hat{\boldsymbol{\psi}}^{(r-1)})$ and $Q_m^{(1)}$ and $Q_m^{(2)}$ are the vector and matrix of first and second derivatives of Q_m respectively. A sandwich estimate of the variance

is obtained by substituting $\hat{\psi}^{(r)}$ in place of $\hat{\psi}^{*(r)}$ and using the estimate

$$\hat{var}\{Q_m^{(1)}(\hat{\psi}^{*(r)}|\hat{\psi}^{(r-1)})\} = \frac{1}{m^2}\sum_{k=1}^{m}\{\frac{\partial lnf(\mathbf{y},\mathbf{b}^{(k)};\hat{\psi}^{(r)})}{\partial\psi}\}\{\frac{\partial lnf(\mathbf{y},\mathbf{b}^{(k)};\hat{\psi}^{(r)})}{\partial\psi^T}\}.$$

In summary, the algorithm works as follows:

1. Select initial estimate $\hat{\psi}^{(0)}$. Set $r = 0$.

2. Increase r by 1.

E-step: For each subject i, $i = 1,...n$ generate m random samples from the distribution of $\mathbf{b}_i|\mathbf{y}_i;\hat{\psi}^{(r-1)}$ using rejection sampling.

3. **M-step:** Update the parameter estimate of ψ by maximizing (3.2), (3.3) and (3.4).

4. Iterate between (2) and (3) until convergence is achieved.

Standard Errors Estimation

After the algorithm has converged standard errors can be estimated using Louis' method of approximation of the observed information matrix (Tanner, 1993). Denote the observed data log-likelihood by l and then the observed data information matrix is

$$-\frac{\partial^2 l(\mathbf{y},\psi)}{\partial\psi\partial\psi^T} = -E[\frac{\partial^2 lnL_u(\mathbf{b},\mathbf{y},\psi)}{\partial\psi\partial\psi^T}|\mathbf{y}] - Var[\frac{\partial lnL_u(\mathbf{b},\mathbf{y},\psi)}{\partial\psi}|\mathbf{y}] =$$

$$= -E[\frac{\partial^2 lnL_u(\mathbf{b},\mathbf{y},\psi)}{\partial\psi\partial\psi^T} + \frac{\partial lnL_u(\mathbf{b},\mathbf{y},\psi)}{\partial\psi}\frac{\partial lnL_u(\mathbf{b},\mathbf{y},\psi)}{\partial\psi^T}|\mathbf{y}]$$

$$+E[\frac{\partial lnL_u(\mathbf{b},\mathbf{y},\psi)}{\partial\psi}|\mathbf{y}]E[\frac{\partial lnL_u(\mathbf{b},\mathbf{y},\psi)}{\partial\psi^T}|\mathbf{y}].$$

By simulating values $\mathbf{b}_i^{(k)}$, $k = 1,...m$, $i = 1,...n$ from the conditional distributions of $\mathbf{b}_i|\mathbf{y}_i$ at the final $\hat{\psi}$ the conditional expectations above can be approximated by the

Monte Carlo sums

$$\frac{1}{m}\sum_{k=1}^{m}\{\frac{\partial^2 lnL_u(\mathbf{b}^{(k)},\mathbf{y},\hat{\psi})}{\partial\psi\partial\psi^T}+\frac{\partial lnL_u(\mathbf{b}^{(k)},\mathbf{y},\hat{\psi})}{\partial\psi}\frac{\partial lnL_u(\mathbf{b}^{(k)},\mathbf{y},\hat{\psi})}{\partial\psi^T}\}$$

and

$$\frac{1}{m}\sum_{k=1}^{m}\frac{\partial lnL_u(\mathbf{b}^{(k)},\mathbf{y},\hat{\psi})}{\partial\psi}.$$

In the above expressions the argument $\hat{\psi}$ means that the derivatives are evaluated at the final parameter estimate. Exactly as in the Gauss-Hermite quadrature example $\hat{Var}(\hat{\psi})=(-\frac{\partial^2 l}{\partial\psi\partial\psi^T}\big|_{\hat{\psi}})^{-1}$.

As the simulation sample size increases the Monte Carlo EM algorithm behaves as a deterministic EM algorithm and usually converges to the exact maximum likelihood estimates, but it may also converge to a local instead of a global maximum (Wu, 1983). A drawback of the Monte Carlo EM algorithm is that it may be very computationally intensive. Hence important alternatives to this 'exact' method for GLMM are methods based on analytical approximations to the marginal likelihood such as the procedures suggested by Breslow and Clayton (1993) and by Wolfinger and O'Connell (1993). Wolfinger and O'Connell propose finding approximate maximum likelihood estimates by approximating the non-normal responses using Taylor series expansions and then solving the linear mixed model equations for the associated normal model. The extension to their method is presented in the next section.

3.2.3 Pseudo-likelihood Approach

Parameter Estimation

Let us first assume for simplicity that the shape parameters ϕ_1 and ϕ_2 for both response distributions are equal to 1. By the model definition

$$\boldsymbol{\mu}_{i1}=g_1^{-1}(\mathbf{X}_{i1}\boldsymbol{\beta}_1+\mathbf{Z}_{i1}\mathbf{b}_{i1})$$

$$\boldsymbol{\mu}_{i2} = g_2^{-1}(\mathbf{X}_{i2}\boldsymbol{\beta}_2 + \mathbf{Z}_{i2}\mathbf{b}_{i2}),$$

where g_1^{-1} and g_2^{-1} above are evaluated componentwise at the elements of $\mathbf{X}_{i1}\boldsymbol{\beta}_1 + \mathbf{Z}_{i1}\mathbf{b}_{i1}$ and $\mathbf{X}_{i2}\boldsymbol{\beta}_2 + \mathbf{Z}_{i2}\mathbf{b}_{i2}$ respectively. Let $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\beta}}_2$, $\hat{\mathbf{b}}_{i1}$ and $\hat{\mathbf{b}}_{i2}$, $i = 1, ...n$, be some estimates of the fixed and the random effects. The corresponding mean estimates are denoted by $\hat{\boldsymbol{\mu}}_{i1}$ and $\hat{\boldsymbol{\mu}}_{i2}$. In a neighborhood around the fixed and random effects estimates the errors $\boldsymbol{\epsilon}_{i1} = \mathbf{y}_{i1} - \boldsymbol{\mu}_{i1}$ and $\boldsymbol{\epsilon}_{i2} = \mathbf{y}_{i2} - \boldsymbol{\mu}_{i2}$ are then approximated by first order Taylor series expansions. Denote the two approximations by

$$\hat{\boldsymbol{\epsilon}}_{i1} = \mathbf{y}_{i1} - \hat{\boldsymbol{\mu}}_{i1} - (g_1^{-1})'(\mathbf{X}_{i1}\hat{\boldsymbol{\beta}}_1 + \mathbf{Z}_{i1}\hat{\mathbf{b}}_{i1})(\mathbf{X}_{i1}\boldsymbol{\beta}_1 + \mathbf{Z}_{i1}\mathbf{b}_{i1} - \mathbf{X}_{i1}\hat{\boldsymbol{\beta}}_1 - \mathbf{Z}_{i1}\hat{\mathbf{b}}_{i1})$$

and

$$\hat{\boldsymbol{\epsilon}}_{i2} = \mathbf{y}_{i2} - \hat{\boldsymbol{\mu}}_{i2} - (g_2^{-1})'(\mathbf{X}_{i2}\hat{\boldsymbol{\beta}}_2 + \mathbf{Z}_{i2}\hat{\mathbf{b}}_{i2})(\mathbf{X}_{i2}\boldsymbol{\beta}_2 + \mathbf{Z}_{i2}\mathbf{b}_{i2} - \mathbf{X}_{i2}\hat{\boldsymbol{\beta}}_2 - \mathbf{Z}_{i2}\hat{\mathbf{b}}_{i2}),$$

where $(g_1^{-1})'(\mathbf{X}_{i1}\hat{\boldsymbol{\beta}}_1 + \mathbf{Z}_{i1}\hat{\mathbf{b}}_{i1})$ and $(g_2^{-1})'(\mathbf{X}_{i2}\hat{\boldsymbol{\beta}}_2 + \mathbf{Z}_{i2}\hat{\mathbf{b}}_{i2})$ are diagonal matrices with elements consisting of evaluations of the first derivatives of g_1^{-1} and g_2^{-1}. Note that because the estimates of the random effects are not consistent as the number of subjects increases this expansion may not work well when the number of observations per subject is small.

Now as in Wolfinger and O'Connell we approximate the conditional distributions of $\boldsymbol{\epsilon}_{i1}|\mathbf{b}_i$ and $\boldsymbol{\epsilon}_{i2}|\mathbf{b}_i$ by normal distributions with $\mathbf{0}$ means and diagonal variance matrices $\mathbf{V}_1(\boldsymbol{\mu}_{i1}) = diag(V_1(\mu_{i11}), ...V_1(\mu_{i1n_i}))$ and $V_2(\boldsymbol{\mu}_{i2}) = diag(V_2(\mu_{i21}), ...V_2(\mu_{i2n_i}))$ respectively, where $V_1(\cdot)$ and $V_2(\cdot)$ are the variance functions of the original responses. Note that it is reasonable to assume that the two normal distributions are uncorrelated because the responses are assumed to be conditionally independent given the random effects. The normal approximation, on the other hand, may not be appropriate for some distributions, such as the Bernoulli distribution.

Substituting $\hat{\boldsymbol{\mu}}_{i1}$ and $\hat{\boldsymbol{\mu}}_{i2}$ in the variance expressions, and using the fact that $(g_1^{-1})'(\mathbf{X}_{i1}\hat{\boldsymbol{\beta}}_1 + \mathbf{Z}_{i1}\hat{\mathbf{b}}_{i1}) = (g_1'(\hat{\boldsymbol{\mu}_{i1}}))^{-1}$ and $(g_2^{-1})'(\mathbf{X}_{i2}\hat{\boldsymbol{\beta}}_2 + \mathbf{Z}_{i2}\hat{\mathbf{b}}_{i2}) = (g_2'(\hat{\boldsymbol{\mu}_{i2}}))^{-1}$, the

conditional distributions of $g_1'(\hat{\boldsymbol{\mu}}_{i1})(\mathbf{y}_{i1} - \hat{\boldsymbol{\mu}}_{i1})$ and $g_2'(\hat{\boldsymbol{\mu}}_{i2})(\mathbf{y}_{i2} - \hat{\boldsymbol{\mu}}_{i2})$ are also multivariate normal with appropriately transformed variance-covariance matrices. Defining $\boldsymbol{\nu}_{i1} = g_1(\hat{\boldsymbol{\mu}}_{i1}) + g_1'(\hat{\boldsymbol{\mu}}_{i1})(\mathbf{y}_{i1} - \hat{\boldsymbol{\mu}}_{i1})$ and $\boldsymbol{\nu}_{i2} = g_2(\hat{\boldsymbol{\mu}}_{i2}) + g_2'(\hat{\boldsymbol{\mu}}_{i2})(\mathbf{y}_{i2} - \hat{\boldsymbol{\mu}}_{i2})$, it follows that

$$\boldsymbol{\nu}_{i1}|\mathbf{b}_i \sim N_{n_i}(\mathbf{X}_{i1}\boldsymbol{\beta}_1 + \mathbf{Z}_{i1}\mathbf{b}_{i1}, g_1'(\hat{\boldsymbol{\mu}}_{i1})\mathbf{V}_1(\hat{\boldsymbol{\mu}}_{i1})g_1'(\hat{\boldsymbol{\mu}}_{i1}))$$

$$\boldsymbol{\nu}_{i2}|\mathbf{b}_i \sim N_{n_i}(\mathbf{X}_{i2}\boldsymbol{\beta}_2 + \mathbf{Z}_{i2}\mathbf{b}_{i2}, g_2'(\hat{\boldsymbol{\mu}}_{i2})\mathbf{V}_2(\hat{\boldsymbol{\mu}}_{i2})g_2'(\hat{\boldsymbol{\mu}}_{i2})).$$

Now, recalling that $\mathbf{b}_i$ has a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, this approximation takes form of a linear mixed model with response $Vec(\boldsymbol{\nu}_{i1}, \boldsymbol{\nu}_{i2})$, random effect $Vec(\mathbf{b}_{i1}, \mathbf{b}_{i2})$, and uncorrelated but heteroscedastic errors. Such a model is called a weighted linear mixed model with a diagonal weight matrix $\hat{\mathbf{W}} = diag(\hat{\mathbf{W}}_i)$, where $\mathbf{W}_i = diag(\mathbf{W}_{i1}, \mathbf{W}_{i2})$, $\hat{\mathbf{W}}_{i1}^{-1} = g_1'(\hat{\boldsymbol{\mu}}_{i1})\mathbf{V}_1(\hat{\boldsymbol{\mu}}_{i1})g_1'(\hat{\boldsymbol{\mu}}_{i1})$ and $\hat{\mathbf{W}}_{i2}^{-1} = g_2'(\hat{\boldsymbol{\mu}}_{i2})\mathbf{V}_2(\hat{\boldsymbol{\mu}}_{i2})g_2'(\hat{\boldsymbol{\mu}}_{i2})$. For canonical link functions $\mathbf{V}_1(\hat{\boldsymbol{\mu}}_{i1}) = [g_1'(\hat{\boldsymbol{\mu}}_{i1})]^{-1}$ and $\mathbf{V}_2(\hat{\boldsymbol{\mu}}_{i2}) = [g_2'(\hat{\boldsymbol{\mu}}_{i2})]^{-1}$, and then $\mathbf{W}_i^{-1} = diag(g_1'(\hat{\boldsymbol{\mu}}_{i1}), g_2'(\hat{\boldsymbol{\mu}}_{i2}))$.

Estimates of $\boldsymbol{\beta} = Vec(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $\mathbf{b}_i$ can be obtained by solving the mixed-model equations below (Harville, 1977):

$$\begin{pmatrix} \mathbf{X}^T\hat{\mathbf{W}}\mathbf{X} & \mathbf{X}^T\hat{\mathbf{W}}\mathbf{Z}\boldsymbol{\Sigma} \\ \mathbf{Z}^T\hat{\mathbf{W}}\mathbf{X} & \boldsymbol{\Sigma} + \mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\hat{\mathbf{W}}\boldsymbol{\nu} \\ \mathbf{Z}^T\hat{\mathbf{W}}\boldsymbol{\nu} \end{pmatrix} \tag{3.5}$$

Here

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \cdot \\ \vdots \\ \mathbf{X}_n \end{pmatrix}, \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 \\ \cdot \\ \vdots \\ \mathbf{Z}_n \end{pmatrix}, \boldsymbol{\nu} = \begin{pmatrix} \boldsymbol{\nu}_1 \\ \cdot \\ \vdots \\ \boldsymbol{\nu}_n \end{pmatrix}, \mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \cdot \\ \vdots \\ \mathbf{b}_n \end{pmatrix}$$

and

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{X}_{i1} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_{i2} \end{pmatrix}, \mathbf{Z}_i = \begin{pmatrix} \mathbf{Z}_{i1} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{i2}, \end{pmatrix}, \boldsymbol{\nu}_i = \begin{pmatrix} \boldsymbol{\nu}_{i1} \\ \boldsymbol{\nu}_{i2} \end{pmatrix}.$$

Variance component estimates can be obtained by numerically maximizing the weighted linear mixed model log-likelihood

$$l = -\frac{1}{2}\sum_{i=1}^{n} ln|\mathbf{V}_i| - \frac{1}{2}\sum_{i=1}^{n} \mathbf{r}_i^T \mathbf{V}_i^{-1} \mathbf{r}_i, \tag{3.6}$$

where

$$\mathbf{V}_i = \mathbf{W}_i^{-1} + \mathbf{Z}_i \boldsymbol{\Sigma} \mathbf{Z}_i^T$$

and

$$\mathbf{r}_i = \boldsymbol{\nu}_i - \mathbf{X}_i(\mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i)^{-1} \mathbf{X}_i^T \mathbf{V}_i^{-1} \boldsymbol{\nu}_i.$$

The algorithm then is as follows:

1. Obtain initial estimates $\hat{\boldsymbol{\mu}}_{i1}^{(0)}$ and $\hat{\boldsymbol{\mu}}_{i2}^{(0)}$ using the original data, $i = 1, ...n$.

2. Compute the modified responses

$\boldsymbol{\nu}_{i1} = g_1(\hat{\boldsymbol{\mu}}_{i1}) + (\mathbf{y}_{i1} - \hat{\boldsymbol{\mu}}_{i1}) g_1'(\hat{\boldsymbol{\mu}}_{i1})$

and

$\boldsymbol{\nu}_{i2} = g_2(\hat{\boldsymbol{\mu}}_{i2}) + (\mathbf{y}_{i2} - \hat{\boldsymbol{\mu}}_{i2}) g_2'(\hat{\boldsymbol{\mu}}_{i2})$.

3. Maximize (3.6) with respect to the variance components.

Stop if the difference between the new and the old variance component estimates is sufficiently small. Otherwise go to the next step.

4. Compute the mixed-model estimates for $\boldsymbol{\beta}$ and $\mathbf{b}_i$ by solving the mixed model equations (3.5):

$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{n} \mathbf{X}_i^T \hat{\mathbf{V}}_i^{-1} \mathbf{X}_i)^{-1} (\sum_{i=1}^{n} \mathbf{X}_i^T \hat{\mathbf{V}}_i^{-1} \boldsymbol{\nu}_i)$,

$\hat{\mathbf{b}}_i = \hat{\boldsymbol{\Sigma}} \mathbf{Z}_i^T \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{r}}_i$.

5. Compute the new estimates of $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{i1}, \boldsymbol{\mu}_{i2})$:

$\hat{\boldsymbol{\mu}}_1 = g_1^{-1}(\mathbf{X}_{i1} \hat{\boldsymbol{\beta}}_1 + \mathbf{Z}_{i1} \hat{\mathbf{b}}_{i1})$

and

$\hat{\boldsymbol{\mu}}_2 = g_2^{-1}(\mathbf{X}_{i2} \hat{\boldsymbol{\beta}}_2 + \mathbf{Z}_{i2} \hat{\mathbf{b}}_{i2})$.

Go to step 2.

In comparison to the original Wolfinger and O'Connell algorithm, the above algorithm for multiple responses is computationally more intensive at step 3 because of the more complicated structure of the variance-covariance matrices $\mathbf{V}_i$. As the number of response variables increases the computational advantages of this approximate procedure over the "exact" maximum likelihood procedures may become less pronounced.

The Newton-Raphson method is a logical choice for the numerical maximization in step 3 because of its relatively fast convergence and because as a side product it provides the Hessian matrix for the approximation of the standard errors of the variance components.

In the general case when ϕ_1 and ϕ_2 are arbitrary, the variance functions $\mathbf{V}_1(\hat{\boldsymbol{\mu}}_{i1})$ and $\mathbf{V}_2(\hat{\boldsymbol{\mu}}_{i2})$ are replaced by $\phi_1\mathbf{V}_1(\hat{\boldsymbol{\mu}}_{i1})$ and $\phi_2\mathbf{V}_2(\hat{\boldsymbol{\mu}}_{i2})$, and hence the weight matrices are accordingly modified: $\hat{\mathbf{W}}_{i1}^{-1} = \hat{\phi}_1 g_1'(\hat{\boldsymbol{\mu}}_{i1})\mathbf{V}_1(\hat{\boldsymbol{\mu}}_{i1})g_1'(\hat{\boldsymbol{\mu}}_{i1})$ and $\hat{\mathbf{W}}_{i2}^{-1} = \hat{\phi}_2 g_2'(\hat{\boldsymbol{\mu}}_{i2})\mathbf{V}_2(\hat{\boldsymbol{\mu}}_{i2})g_2'(\hat{\boldsymbol{\mu}}_{i2})$. The estimates of ϕ_1 and ϕ_2 can be obtained in step 3 together with the other variance components estimates. This approach is different from the approach taken by Wolfinger and O'Connell, who estimated the dispersion parameter for the response together with the fixed and random effects.

Standard Error Estimation

The standard errors for the regression parameter and for the random effects estimates are approximated from the linear mixed model:

$$Var(\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}) \approx \begin{pmatrix} \mathbf{X}^T\hat{\mathbf{W}}\mathbf{X} & \mathbf{X}^T\hat{\mathbf{W}}\mathbf{Z}\hat{\boldsymbol{\Sigma}} \\ \mathbf{Z}^T\hat{\mathbf{W}}\mathbf{X} & \hat{\boldsymbol{\Sigma}} + \mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z} \end{pmatrix}^{-1}$$

As $\boldsymbol{\Sigma}$ and $\mathbf{W}$ are unknown, estimates of the variance components will be used.

The standard errors for the variance components can be approximated by using exact or numerical derivatives of the likelihood function in step 3 of the algorithm.

For the same reasons as discussed in Section 3.2 it is better to maximize with respect to the Choleski factors of $\boldsymbol{\Sigma}$. The computation of the standard errors, however, should be carried out with respect to the original parameters σ_1, σ_2 and ρ (or σ_1^2, σ_2^2 and σ_{12}). This can be done only if the final variance-covariance estimate is positive definite. The approach discussed in Section 3.2 for representation of the variance-covariance matrix can also be adopted here to simplify computations.

3.3 Simulated Data Example

The methods discussed in the previous sections are illustrated on one simulated data set in which the true values on the parameters are known. The data consists of $J = 10$ repeated measures of a normal and of a Bernoulli response on each of $I = 30$ subjects. No covariates are considered and a random intercept is assumed for each variable. In the notation introduced in Chapter 2,

$$y_{i1j}|b_{i1} \sim \text{indep. } N(\mu_{i1j}, \sigma^2),$$

$$y_{i2j}|b_{i2} \sim \text{indep. } Bernoulli(\mu_{i2j}),$$

$$\mu_{i1j} = \beta_1 + b_{i1},$$

$$logit(\mu_{i2j}) = \beta_2 + b_{i2},$$

$$\mathbf{b}_i = (b_{i1}, b_{i2})^T \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}),\ \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

This means that we define a normal theory random intercept model for one of the responses, a logistic regression model with a random intercept for the other response, and we assume that the random intercepts are correlated. The data are simulated for $\sigma^2 = 1$, $\beta_1 = 4$, $\beta_2 = 1$, $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$. The parameter estimates and their standard errors for the three fitting methods discussed above are presented in Table 3.1. The methods are programmed in Ox and are run on Sun Ultra workstations. Ox is an object-oriented matrix programming language developed at Oxford University

Table 3.1. Estimates and standard errors from model fitting for the simulated data example using Gauss-Hermite quadrature (GQ), Monte Carlo EM algorithm (MCEM) and pseudo-likelihood (PL)

Parameter	True	GQ		MCEM		PL	
	value	Est.	SE	Est.	SE	Est.	SE
β_1	4.0	3.72	0.19	3.72	0.20	3.72	0.19
β_2	1.0	1.08	0.18	1.08	0.19	1.08	0.17
σ^2	1.0	0.97	0.08	0.97	0.08	0.97	0.08
σ_1^2	1.0	0.94	0.27	0.94	0.27	0.94	0.27
σ_2^2	0.5	0.39	0.26	0.41	0.35	0.35	0.22
σ_{12}	0.5	0.54	0.21	0.54	0.24	0.48	0.20

(Doornik, 1998). It is convenient because of the availability of matrix operations and a variety of predefined functions and even more importantly because of the speed of computations.

All standard errors, except those for the Monte Carlo EM algorithm and those for the regression parameters for the pseudo-likelihood algorithm, are based on numerical derivatives. The identity matrix is used as an initial estimate of the covariance matrix for the random components, and the initial estimates for β_1, β_2 and σ^2 are taken to be the sample means for the normal and the Bernoulli response, and the sample variance for the normal response. Gauss-Hermite quadrature is used with number of quadrature points in each dimension varying from 10 to 100.

The $MaxBFGS$ numerical maximization procedure in Ox is used for optimization in the Gaussian quadrature and the pseudo-likelihood algorithms with a convergence criterion based on $\delta_1 = 0.0001$. The convergence criterion used for the Monte Carlo EM algorithm is based on $\delta_1 = 0.003$. $MaxBFGS$ employs a quasi-Newton method for maximization developed by Broyden, Fletcher, Goldfarb and Shanno (BFGS) (Doornik, 1998) and it can use either analytical or numerical first derivatives.

The times to achieve convergence by the three methods are 30 minutes for Gaussian quadrature (GQ) with 100 quadrature points in each dimension, approximately 3 hours for the Monte Carlo EM algorithm and 1.5 hours for the pseudo-likelihood

(PL) algorithm. The numbers of iterations required are 22, 53 and 7 respectively. The final simulation sample size for the Monte Carlo EM algorithm is 9864.

As expected, the Monte Carlo EM estimates are very close to the Gaussian quadrature estimates with comparable standard errors. Only the estimate of σ_2^2 is slightly different and has a rather large standard error, but this is due to premature stopping of the EM algorithm. This problem can be solved by increasing the required precision for convergence but at the expense of a very large simulation sample size.

As can be seen from the table, the parameters corresponding to the normal response are well estimated by the pseudo-likelihood algorithm, but some of the estimates corresponding to the Bernoulli response are somewhat underestimated and have smaller standard errors. The estimate of the random intercept variance σ_2^2 is about 10% smaller than the corresponding Gaussian quadrature estimate with about 15% smaller standard error. The estimate of σ_{12} is also about 15% smaller than the corresponding Gauss-Hermite quadrature and Monte Carlo estimates, but it is actually closer to the true value in this particular example. In general the variance component estimates for binary data obtained using the pseudo-likelihood algorithm are expected to be biased downward as indicated by previous research in the univariate GLMM (Breslow and Clayton, 1993; Wolfinger, 1998). Hence the Wolfinger and O'Connell method should be used with caution for the extended GLMM's with at least one Bernoulli response, keeping in mind the possible attenuation of estimates and of the standard errors.

Gauss-Hermite Quadrature: 'Exact' vs 'Numerical' Derivatives

As mentioned before $MaxBFGS$ can be used either with exact or with numerical first derivatives. Comparisons of these two approaches in terms of computational speed and accuracy are provided in Tables 3.2 and 3.3 respectively. The 'Iteration' columns in Table 3.2 contain number of iterations and the 'Convergence' columns

Table 3.2. 'Exact' versus 'numerical' derivatives for the simulated data example: convergence using Gauss-Hermite quadrature.

Number of quad. points	Exact derivatives Time (min.)	Iteration	Convergence	Numerical derivatives Time (min.)	Iteration	Convergence
10	27.4	56	no	0.3	19	strong
20	21.3	22	weak	1.1	19	strong
30	75	35	weak	2.8	23	strong
40	120	34	strong	4.8	22	strong
50	246	47	strong	7.2	22	strong
60	237	34	strong	10.9	22	strong

show whether the convergence tests are passed at the prespecified δ_1 level (strong convergence), or passed at the lower level $\delta_1 = 0.001$ (weak convergence) or not passed at all (no convergence). Note that for the exact derivatives strong convergence is achieved with 40 or more quadrature points.

It is rather surprising that the 'exact' Gaussian quadrature procedure does not perform better than the 'numerical' one. The parameter estimates from both procedures are the same (Table 3.3) for 40 or more quadrature points but the 'exact' Gauss-Hermite quadrature converges much more slowly than the numerical procedure (hours as compared to minutes). The numerical procedure also has the advantage of simpler programming code and hence for the rest of the disertation will be the preferred quadrature method.

Monte Carlo EM algorithm: Variability of Estimated Standard Errors

We observed that the estimated Monte Carlo EM standard errors varied somewhat for different random seeds, so we generated 100 samples with the final simulation sample size and computed the average and the standard deviation of the standard error estimates for these samples. Table 3.4 shows the results. We include the Gaussian quadrature standard error estimates for comparison. There is much variability in the standard error estimates (especially in the estimate of σ_2^2) which is primarily due to

Table 3.3. 'Exact' versus 'numerical' derivatives for the simulated data example: estimates and standard errors using Gauss-Hermite quadrature with varrying number of quadrature points (m).

m	Exact derivatives						Numerical derivatives					
	β_1	β_2	σ^2	σ_1^2	σ_2^2	σ_{12}	β_1	β_2	σ^2	σ_1^2	σ_2^2	σ_{12}
	SE	SE	SE	SE	SE	SE	SE	SE	SE	SE	SE	SE
10	3.68	1.06	0.97	0.98	0.38	0.52	3.72	1.08	0.97	0.85	0.35	0.48
	0.13	0.16	0.08	0.28	0.25	0.22	0.12	0.16	0.08	0.16	0.24	0.15
20	3.75	1.10	0.96	0.95	0.39	0.54	3.75	1.10	0.97	0.95	0.39	0.54
	0.16	0.18	0.08	0.27	0.26	0.21	0.16	0.18	0.08	0.27	0.26	0.21
30	3.71	1.07	0.97	0.94	0.39	0.54	3.71	1.07	0.97	0.95	0.39	0.54
	0.19	0.18	0.08	0.27	0.26	0.21	0.19	0.18	0.08	0.29	0.27	0.22
40	3.72	1.08	0.97	0.94	0.39	0.54	3.72	1.08	0.97	0.94	0.39	0.54
	0.18	0.18	0.08	0.27	0.26	0.21	0.18	0.18	0.08	0.26	0.26	0.20
50	3.72	1.08	0.97	0.94	0.39	0.54	3.72	1.08	0.97	0.94	0.39	0.54
	0.19	0.18	0.08	0.27	0.26	0.21	0.19	0.18	0.08	0.27	0.26	0.21
60	3.72	1.08	0.97	0.94	0.39	0.54	3.72	1.08	0.97	0.94	0.39	0.54
	0.19	0.18	0.08	0.27	0.26	0.21	0.19	0.18	0.08	0.27	0.26	0.21

several extreme observations. If the simulation sample size is increased by one-third, the standard error estimates are more stable (see the last two columns of Table 3.4). It is not surprising that there is more variability in the standard error estimates than in the parameter estimates because only the latter are controlled by the convergence criterion. So it may be necessary to increase the simulation sample size to obtain better estimates of the standard errors. This, however, requires additional computational effort and it is not clear by how much the simulation sample size should be increased. We consider a different approach to dealing with standard error instability in the next section of the dissertation.

Comparison of Simultaneous and Separate Fitting of the Response Variables

To investigate the possible efficiency gains of joint fitting of the two responses over separate fitting, we compared estimates from the joint to estimates from the two separate fits. The results for the three estimation methods are provided in Table

Table 3.4. Variability in Monte Carlo EM standard errors for the simulated data example. Means and standard deviations of the standard error estimates are computed for 100 samples using two different simulation sample sizes. The Gauss-Hermite quadrature standard errors are given as a reference in the second column.

Parameter	GQ SE	Mean SE	SD of SE	Mean SE	SD of SE
		ss=9864		ss=13152	
β_1	0.19	0.21	0.07	0.19	0.03
β_2	0.18	0.20	0.09	0.19	0.03
σ^2	0.08	0.08	0.0008	0.08	0.0001
σ_1^2	0.27	0.27	0.03	0.27	0.01
σ_2^2	0.26	0.41	0.76	0.31	0.13
σ_{12}	0.21	0.22	0.04	0.22	0.02

Table 3.5. Results from joint and separate Gauss-Hermite quadrature of the two response variables in the simulated data example

Parameter	True value	Joint estimation		Separate estimation	
		Estimate	SE	Estimate	SE
β_1	4.0	3.72	0.19	3.72	0.19
β_2	1.0	1.08	0.18	1.07	0.18
σ^2	1.0	0.97	0.08	0.97	0.08
σ_1^2	1.0	0.94	0.27	0.94	0.27
σ_2^2	0.5	0.39	0.26	0.37	0.25
σ_{12}	0.5	0.54	0.21	–	–

3.5 (Gaussian quadrature), Table 3.6 (Monte Carlo EM), and Table 3.7 (pseudo-likelihood). For the separate fits we used PROC NLMIXED for Gaussian quadrature and wrote our own programs in Ox for Monte Carlo EM and pseudo-likelihood. Notice that when the Normal response is considered separately, the corresponding model is a one-way random effects ANOVA and can be fitted using PROC MIXED in SAS, for example. The estimates from PROC MIXED are as follows: $\hat{\beta}_1 = 3.72$ ($SE = 0.19$), $\hat{\sigma}^2 = 0.97$ ($SE = 0.08$) and $\hat{\sigma}_1^2 = 0.94$ ($SE = 0.27$).

It is surprising that although the estimated correlation between the random effects was rather high ($\hat{\rho} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = 0.86$) the estimates and the estimated standard errors from the joint and from the separate fits were very similar. Having in mind that it is much faster to fit the responses separately and that there is existing software for univariate repeated measures, it is not clear if there are any advantages to joint fitting

Table 3.6. Results from joint and separate Monte Carlo EM estimation of the two response variables in the simulated data example

Parameter	True value	Joint estimation		Separate estimation	
		Estimate	SE	Estimate	SE
β_1	4.0	3.72	0.20	3.73	0.22
β_2	1.0	1.08	0.19	1.08	0.18
σ^2	1.0	0.97	0.08	0.97	0.08
σ_1^2	1.0	0.94	0.27	0.95	0.28
σ_2^2	0.5	0.41	0.35	0.38	0.23
σ_{12}	0.5	0.54	0.24	–	–

Table 3.7. Results from joint and separate pseudo-likelihood estimation of the two response variables in the simulated data example

Parameter	True value	Joint estimation		Separate estimation	
		Estimate	SE	Estimate	SE
β_1	4.0	3.72	0.19	3.72	0.19
β_2	1.0	1.08	0.17	1.05	0.17
σ^2	1.0	0.97	0.08	0.97	0.08
σ_1^2	1.0	0.94	0.27	0.94	0.27
σ_2^2	0.5	0.35	0.22	0.35	0.22
σ_{12}	0.5	0.48	0.20	–	–

in this particular example. Advantages and disadvantages of joint and separate fitting will be discussed in Chapter 5 of the dissertation.

3.4 Applications

3.4.1 Developmental Toxicity Study in Mice

In this section the 'exact' maximum likelihood methods are applied to the ethylene glycol (EG) example mentioned in the introduction. The model is defined as follows (d_i denotes the exposure level of EG for the i^{th} dam):

y_{i1j} - birth weight of j^{th} live fetus in i^{th} litter.

y_{i2j} - malformation status of j^{th} live fetus in i^{th} litter.

$y_{i1j}|b_{i1} \sim indep.\, N(\mu_{i1j}, \sigma^2)$.

$y_{i2j}|b_{i2} \sim indep.\, Be(\mu_{i2j})$.

Table 3.8. Gauss-Hermite and Monte Carlo EM estimates and standard errors from model fitting with a quadratic effect of dose on birth weight in the ethylene glycol data

Parameter	Gauss-Hermite quadrature		Monte Carlo EM	
	Estimate	SE	Estimate	SE
β_{10}	0.962	0.016	0.964	0.037
β_{11}	-0.118	0.028	-0.120	0.053
β_{12}	0.010	0.009	0.010	0.015
β_{20}	-4.267	0.409	-4.287	0.455
β_{21}	1.720	0.207	1.731	0.217
σ^2	0.006	0.0003	0.006	0.0003
σ_1^2	0.007	0.001	0.007	0.001
σ_2^2	2.287	0.596	2.238	0.560
σ_{12}	-0.082	0.020	-0.082	0.022

Table 3.9. Gauss-Hermite quadrature and Monte Carlo EM estimates and standard errors from model fitting with linear trends of dose in the ethylene glycol data

Parameter	GQ (logit)		MCEM (logit)		GQ (probit)		MCEM (probit)	
	Est	SE	Est	SE	Est	SE	Est	SE
β_{10}	0.952	0.014	0.952	0.021	0.952	0.014	0.954	0.028
β_{11}	-0.087	0.008	-0.088	0.008	-0.087	0.008	-0.088	0.012
β_{20}	-4.335	0.411	-4.335	0.522	-2.340	0.213	-2.422	0.269
β_{21}	1.749	0.208	1.752	0.225	0.970	0.111	0.982	0.129
σ	0.075	0.002	0.075	0.002	0.075	0.002	0.075	0.002
σ_1	0.086	0.007	0.086	0.007	0.086	0.007	0.086	0.007
σ_2	1.513	0.196	1.495	0.207	0.839	0.107	0.831	0.101
ρ	-0.682	0.088	-0.687	0.091	-0.666	0.090	-0.670	0.094

$$\mu_{i1j} = \beta_{10} + \beta_{11} * d_i + b_{i1},$$

$$logit(\mu_{i2j}) = \beta_{20} + \beta_{21} * d_i + b_{i2}$$

$$\mathbf{b}_i = (b_{i1}, b_{i2})' \sim MVN(\mathbf{0}, \mathbf{\Sigma}),\ \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

We considered a linear trend in dose in both linear predictors. Although other authors have found a significant quadratic dose trend for birth weight, we found that to be nonsignificant and hence have not included it in the final model (see Table 3.8). We also considered a probit link.

The Gaussian quadrature and Monte Carlo EM parameter estimates from the model fits are given in Table 3.9. Initial estimates for the regression parameters were

the estimates from the fixed effects models for the two responses, σ was set equal to the estimated standard deviation from the linear mixed model for birth weight, and the identity matrix was taken as an initial estimate for the variance-covariance matrix of the random effects. Gaussian quadrature estimates were obtained using 100 quadrature points in each dimension which took 36 hours for both the logit and the probit links. Adaptive Gaussian quadrature will probably be much faster because it requires a smaller number of quadrature points. The Monte Carlo EM algorithm for the model with logit link ran for 3 hours and 6 minutes, and required 41 iterations and a final sample size of 989 for convergence. The Monte Carlo EM algorithm for the model with a probit link ran for 5 hours and 13 minutes, and required 37 iterations and a final sample size of 1757 for convergence. The convergence precisions were the same as in the simulated data example.

The estimates from the two algorithms are similar but the standard error estimates from the Monte Carlo EM algorithm are larger than their counterparts from Gaussian quadrature. This is not always the case as seen from Table 3.10. There is much variability in the standard error estimates from the Monte Carlo EM algorithm. The estimates can be improved by considering larger simulation sample sizes but we adopt a different approach. The observed information matrix at each step of the algorithm is easily approximated using Louis' method and it does not require extra computational effort, because it relies on approximations needed for the sample size increase decision. Because the EM algorithm converges slowly in close proximity of the estimates and because the convergence criterion must be satisfied in three consecutive iterations, the approximated observed information matrices from the last three iterations are likely to be equally good in approximating the variance of the parameter estimates. Hence if we average the three matrices we are likely to obtain a better estimate of the variance-covariance matrix than if we rely on only one particular iteration.

Table 3.10. Variability in Monte Carlo EM standard errors for the ethylene glycol example. Means and standard deviations of the standard error estimates are computed for 100 samples for the logit and probit models. Gauss-Hermite quadrature standard errors as given as a reference.

Parameter	Logit model			Probit model		
	GQ	Mean of SE	SD of SE	GQ	Mean of SE	SD of SE
β_{10}	0.014	0.015	0.032	0.014	0.014	0.021
β_{11}	0.008	0.008	0.015	0.008	0.007	0.010
β_{20}	0.411	0.440	0.500	0.213	0.200	0.137
β_{21}	0.208	0.206	0.207	0.111	0.101	0.063
σ	0.002	0.002	0.0001	0.002	0.002	0.00002
σ_1	0.007	0.007	0.0005	0.007	0.007	0.0001
σ_2	0.196	0.212	0.115	0.107	0.110	0.020
ρ	0.088	0.094	0.052	0.090	0.092	0.012

Tables 3.11 and 3.12 give standard error estimates for the EG data based on Gaussian quadrature (column GQ), based on the the Monte Carlo approximated observed information matrix in the third to last (column A1), second to last (column A2) and last (column A3) iterations, and based on average of the observed information matrices of the last two (column A4) and of the last three (column A5) iterations. Clearly using only the estimate of the observed information matrix from the last iteration is not satisfactory because it depends heavily on the random seed and may contain some negative standard error estimates. Averaging over the three final iterations is better, although it is not guaranteed that even the pooled estimate of the observed information matrix will be positive definite, or that it will lead to improved estimates.

All parameter estimates in the model are significantly different from zero with birth weight significantly decreasing with increasing dose and probability for malformation significantly increasing with increasing dose. As expected, the regression parameters estimates using the logit link function are greater than those obtained using the probit link function. The use of the logit link function facilitates the interpretation of the parameter estimates. Thus, if the EG dose level for a litter in increased by $1g/kg$ the estimated odds of malformation for any fetus within that litter increase

Table 3.11. Monte Carlo EM standard errors using logit link in the ethylene glycol example. The approximations are based on the estimate of the observed information matrix from the third to last iteration (A1), second to last iteration (A2), last iteration (A3), the average of the last two iterations (A4), and the average of the last three iterations (A5). The standard error estimates obtained using Gauss-Hermite quadrature are given as a reference in the column labeled GQ.

Parameter	Estimated S.E.					
	GQ	A1	A2	A3	A4	A5
β_{10}	0.014	0.017	0.037	0.012	0.016	0.016
β_{11}	0.008	0.007	0.014	0.001	0.008	0.007
β_{20}	0.411	0.385	0.429	0.365	0.409	0.396
β_{21}	0.208	0.183	0.239	0.215	0.236	0.206
σ	0.002	0.002	0.002	0.002	0.002	0.002
σ_1	0.007	0.007	0.007	0.007	0.007	0.007
σ_2	0.196	0.199	0.202	0.221	0.204	0.199
ρ	0.088	0.081	0.084	0.136	0.096	0.089

Table 3.12. Monte Carlo EM standard errors using probit link in the ethylene glycol example. The approximations are based on the estimate of the observed information matrix from the third to last iteration (A1), second to last iteration (A2), last iteration (A3), the average of the last two iterations (A4), and the average of the last three iterations (A5). The standard error estimates obtained using Gauss-Hermite quadrature are given as a reference in the column labeled GQ.

Parameter	Estimated S.E.					
	GQ	A1	A2	A3	A4	A5
β_{10}	0.014	0.021	0.021	0.006	0.016	0.017
β_{11}	0.008	0.012	0.007	0	0.008	0.009
β_{20}	0.213	0.273	0.227	0	0.243	0.233
β_{21}	0.111	0.151	0.110	0	0.132	0.127
σ	0.002	0.002	0.002	0.002	0.002	0.002
σ_1	0.007	0.008	0.007	0.007	0.007	0.007
σ_2	0.107	0.127	0.097	0	0.124	0.110
ρ	0.090	0.094	0.102	0	0.101	0.095

$exp(1.75) = 5.75$ times. The probit link function does not offer similar easy interpretation, but allows one to obtain population-averaged estimates for dose. If the subject-specific regression estimate is $\hat{\boldsymbol{\beta}}$, then the marginal regression parameters are estimated by $\frac{\hat{\boldsymbol{\beta}}}{\sqrt{(1+\hat{\sigma}^2_{b2})}}$ (Section 2.1). Hence, we obtain a marginal intercept of -1.793 and a marginal slope of 0.743. The latter can be interpreted as the amount by which the population-averaged probability of malformation changes on the probit scale for each unit increase in dose. The subject-specific slope estimate of 0.970 on other hand is interpreted as the increase in the individual fetus probability of malformation on the probit scale for each unit increase in dose. A more meaningful interpretation for those numbers can be offered if one assumes a continuous underlying malformation variable for the observed binary malformation outcome. The latent variable reflects a cumulative detrimental effect which manifests itself in a malformation if it exceeds a certain threshold. The effect of dose on this underlying latent variable is linear and hence $\hat{\beta}_{21}$ is interpreted as the amount by which the cumulative effect is increased for one unit increase in dose. $\hat{\beta}_{21}$ can also be interpreted as the highest rate of change in the probability for malformation (Agresti, 1990, p.103). This highest rate of change is estimated to be achieved at $-\frac{\hat{\beta}_{21}}{\hat{\beta}_{20}}$.

As expected birth weight and malformation are significantly negatively correlated as judged from the Wald test from Table 3.9 : $(\frac{\hat{\rho}}{s.e.(\hat{\rho})})^2 = (\frac{-0.666}{0.090})^2 = 54.8$ ($p-value < 0.0001$). This translates into a negative correlation between the responses but there is no closed form expression to estimate it precisely.

It is interesting to compare the joint and the separate fits of the two response variables (Tables 3.13 and 3.14). Table 3.13 presents estimates obtained using Gaussian quadrature and shows that the estimates for the Normal response are identical (within the reported precision) but the estimates for the Bernoulli response from the separate fits are generally larger in absolute value with larger standard errors.

Table 3.13. Results from joint and separate Gauss-Hermite quadrature of the two response variables in the ethylene glycol example.

Par.	Logit models				Probit models			
	Joint fit		Separate fits		Joint fit		Separate fits	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE
β_{10}	0.952	0.014	0.952	0.014	0.952	0.014	0.952	0.014
β_{11}	-0.087	0.008	-0.087	0.008	-0.087	0.008	-0.087	0.008
β_{20}	-4.335	0.411	-4.356	0.438	-2.340	0.213	-2.426	0.230
β_{21}	1.749	0.208	1.779	0.220	0.970	0.111	0.993	0.118
σ	0.075	0.002	0.075	0.002	0.075	0.002	0.075	0.002
σ_1	0.086	0.007	0.086	0.007	0.086	0.007	0.086	0.007
σ_2	1.513	0.196	1.577	0.211	0.839	0.107	0.881	0.117
ρ	-0.682	0.088	–	–	-0.666	0.090	–	–

Table 3.14. Results from joint and separate fitting of the Monte Carlo EM algorithm for the two response variables in the ethylene glycol example

Par.	Logit models				Probit models			
	Joint fit		Separate fits		Joint fit		Separate fits	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE
β_{10}	0.952	0.021	0.952	0.014	0.954	0.028	0.952	0.014
β_{11}	-0.088	0.008	-0.088	0.007	-0.088	0.012	-0.088	0.007
β_{20}	-4.335	0.522	-4.453	0.344	-2.422	0.269	-2.499	0.193
β_{21}	1.752	0.225	1.822	0.211	0.982	0.129	1.025	0.098
σ	0.075	0.002	0.075	0.002	0.075	0.002	0.075	0.002
σ_1	0.086	0.007	0.086	0.007	0.086	0.007	0.086	0.007
σ_2	1.495	0.207	1.547	0.199	0.831	0.101	0.869	0.052
ρ	-0.687	0.091	–	–	0.670	0.094	–	–

This may indicate small efficiency gains in fitting the responses together rather than separately. More noticeable differences in the parameter estimates and especially in the standard errors are observed in the results from the Monte Carlo EM algorithm (Table 3.14). In this case the standard error estimates for the Bernoulli components from the separate fits are smaller than their counterparts from the joint fits. These results, however, may be deceiving because there is large variability in the standard error estimates depending on the particular simulation sample used (see Table 3.10).

As a result of the conditional independence assumption the correlation between birth weight and malformation within fetus, and the correlation between birth weight and malformation measured on two different fetuses within the same litter, are assumed to be the same. However, in practice this assumption may not be satisfied. We would expect that measurements on the same fetus are more highly correlated than measurements on two different fetuses within a litter. Hence, it is very important to be able to check the conditional independence assumption and to investigate how departures from that assumption affect the parameter estimates. In the following chapter score tests are used to check aspects of conditional independence without fitting more complicated models. When the score tests show that there is non-negligible departure from this assumption, alternative models should be considered. In the case of a binary and a continuous response one more general model is presented in Chapter 5. It has been considered by Catalano and Ryan (1992), who used GEE methods to fit it. We propose "exact" maximum likelihood estimation which allows for direct estimation of the variance-covariance structure.

3.4.2 Myoelectric Activity Study in Ponies

The second data set introduced in Section 2.4 is from a study on myoelectric activity in ponies. The purpose of this analysis is to simultaneously assess the immediate effects of six drugs and placebo on spike burst rate and spike burst duration within

a pony. Two features of this data example are immediately obvious from Table 2.2: there is a large number of observations within cluster (pony) and the variance of the spike burst rate response for each pony is much larger than the mean (see Table 2.2). The implication of the first observation is that ordinary Gauss-Hermite quadrature as described in Section 3.2.1 may not work well because some of the integrands will be essentially zero. Adaptive Gaussian quadrature on the other hand will be more appropriate because it requires a smaller number of quadrature points. Also the additional computations to obtain the necessary modes and curvatures will not slow down the algorithm because there are only 6 subjects in the data set and hence only six additional maximizations will be performed. Adaptive Gaussian quadrature is described in the next subsection of the dissertation.

The implication of the second observation concerns the response distribution of the spike burst rate response. An obvious choice for count data is the Poisson distribution, but the Poisson distribution imposes equality of the mean and the variance of the response and this assumption is clearly not satisfied for the data even within a pony. Therefore some way of accounting for the extra dispersion in addition to the pony effect must be incorporated in the model. Such an approach based on the negative binomial distribution is discussed later in this chapter.

Adaptive Gaussian Quadrature

Adaptive Gaussian quadrature has been considered by Liu and Pierce (1994), Pinheiro and Bates (1999), and Wolfinger (1998). We now present that idea in the context of the bivariate GLMM. Recall that the likelihood for the i^{th} subject has the form

$$L_i = \int_{\mathbf{R}^q} f(\mathbf{y}_i, \mathbf{b}_i) d\mathbf{b}_i,$$

where

$$f(\mathbf{y}_i, \mathbf{b}_i) = \prod_{j=1}^{n_i} \{f_1(y_{i1j}|\mathbf{b}_{i1}; \boldsymbol{\beta}_1, \phi_1) f_2(y_{i2j}|\mathbf{b}_{i2}; \boldsymbol{\beta}_2, \phi_2)\} \frac{1}{|2\pi\boldsymbol{\Sigma}|^{\frac{1}{2}}} exp(-\frac{1}{2}\mathbf{b}_i^T \boldsymbol{\Sigma}^{-1}\mathbf{b}_i).$$

Let $\tilde{\mathbf{b}}_i$ be the mode of $f(\mathbf{y}_i, \mathbf{b}_i)$ and $\tilde{\boldsymbol{\Gamma}}_i = (-\frac{\partial^2 log f(\mathbf{y}_i, \mathbf{b}_i)}{\partial \mathbf{b}_i \partial \mathbf{b}_i^T})^{-1}|_{\tilde{\mathbf{b}}_i}$. Then

$$L_i = \int_{\mathbf{R}^q} f^*(\mathbf{y}_i, \mathbf{b}_i) \phi(\mathbf{b}_i; \tilde{\mathbf{b}}_i, \tilde{\boldsymbol{\Gamma}}_i) d\mathbf{b},$$

where $f^*(\mathbf{y}_i, \mathbf{b}_i) = \frac{f(\mathbf{y}_i, \mathbf{b}_i)}{\phi(\mathbf{b}_i; \tilde{\mathbf{b}}_i, \tilde{\boldsymbol{\Gamma}}_i)}$ and

$$\phi(\mathbf{b}_i; \tilde{\mathbf{b}}_i, \tilde{\boldsymbol{\Gamma}}_i) = exp(-\frac{1}{2}(\mathbf{b}_i - \tilde{\mathbf{b}}_i)^T \tilde{\boldsymbol{\Gamma}}_i^{-1}(\mathbf{b}_i - \tilde{\mathbf{b}}_i)).$$

The needed transformation of $\mathbf{b}_i$ is then $\mathbf{b}_i = \tilde{\mathbf{b}}_i + \sqrt{2}\boldsymbol{\Delta}_i \mathbf{z}_i$, where $\boldsymbol{\Gamma}_i = \boldsymbol{\Delta}_i \boldsymbol{\Delta}_i^T$. Hence each integral is approximated by

$$L_i^{GQ} = |\sqrt{2}\boldsymbol{\Delta}_i| \sum_{k_1=1}^{m} w_{k_1}^{(1)} ... \sum_{k_q=1}^{m} w_{k_q}^{(q)} f^*(\mathbf{z}_i^{(k)})$$

with the tabled univariate weights $w_{k_l}^{(l)}$ and nodes $\mathbf{z}_i^{(k)} = \tilde{\mathbf{b}}_i + \sqrt{2}\boldsymbol{\Delta}_i \mathbf{d}^{(k)}$ for the multiple index $\mathbf{k} = (k_1, ...k_q)$, where $\mathbf{d}^{(k)}$ are the tabled nodes of Gauss-Hermite integration of order m.

The difference between this procedure and the ordinary Gauss-Hermite approximation is in the centering and spread of the nodes that are used. Here they are distributed where most of the data are available and this makes adaptive quadrature more efficient. The computation of the modes $\tilde{\mathbf{b}}_i$ and the curvatures $\tilde{\boldsymbol{\Gamma}}_i$ requires maximization of the integrand function for each subject, which in general requires a numerical procedure, but this was not a major impediment for the pony data. We used $MaxBFGS$ within $MaxBFGS$ to perform the maximization.

Negative Binomial Model

When the dispersion in count data is more than that predicted by the Poison model a convenient parametric approach is to assume a Gamma prior for the Poisson mean (McCullagh and Nelder, 1989). Letting $Y \sim Poisson(\mu)$, then $P(Y = y) = \frac{\mu^y}{y!}e^{-\mu}$, $y = 0, 1, 2, ...$, and $E(Y) = Var(Y) = \mu$. Suppose that we want to specify a larger variance for Y but keep the mean the same. This can be accomplished as suggested by Booth and Hobert (personal communication). Let $Y|u \sim Poisson(u\mu)$ and let $u \sim Gamma(\alpha, \frac{1}{\alpha})$. The density for u is $f(u) = \frac{\alpha^\alpha u^{\alpha-1}e^{-\alpha u}}{\Gamma(\alpha)}$. Then unconditionally Y has a negative binomial distribution with probability density

$$P(Y = y) = \frac{\Gamma(y + \alpha)}{\Gamma(\alpha)y!}(\frac{\alpha}{\alpha + \mu})^\alpha(\frac{\mu}{\alpha + \mu})^y$$

and $E(Y) = \mu$ and $Var(Y) = \mu + \frac{\mu^2}{\alpha}$. Notice that for large α the negative binomial variance approaches the Poisson variance, meaning that is there is little overdispersion, but for small α the negative binomial variance can be much larger than the Poisson variance.

The mean in the above model can be specified as a function of covariates using a log link $ln(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$, $i = 1, ...n$. The log link is convenient because it is the canonical link for the Poisson distribution and transforms the mean so that it can take any real value. It is possible to obtain maximum-likelihood estimates of α and $\boldsymbol{\beta}$ by directly numerically maximizing the negative binomial likelihood but there is a more elegant way based on the EM algorithm as proposed by Booth and Hobert (personal communication).

The complete data consists of $\mathbf{y}$ and $\mathbf{u}$, and the complete data log-likelihood is

$$lnL_V = c + \sum_{i=1}^{n}[\alpha ln(\alpha) - ln\Gamma(\alpha) + y_i ln(\mu_i) + \alpha ln(u_i) - u_i(\mu_i + \alpha)],$$

where c is a constant depending only on the data but not on the unknown parameters. The l^{th} E-step of the EM algorithm then involves calculating

$$E(ln\, L_V(\alpha, \boldsymbol{\beta})|\hat{\alpha}_{(l-1)}, \hat{\boldsymbol{\beta}}_{(l-1)}) = c + N\alpha ln(\alpha) - N\, ln\Gamma(\alpha) +$$

$$\sum_{i=1}^{n}[y_i \mathbf{x}_i^T \boldsymbol{\beta} + \alpha E(ln\, u_i|y_i, \hat{\alpha}_{(l-1)}, \hat{\boldsymbol{\beta}}_{(l-1)}) - E(u_i|y_i, \hat{\alpha}_{(l-1)}, \hat{\boldsymbol{\beta}}_{(l-1)})(exp(\mathbf{x}_i^T \boldsymbol{\beta}) + \alpha)]$$

But because of the choice of the prior distribution for u_i, the posterior distribution for $u_i|y_i$ evaluated at $\hat{\alpha}_{(l-1)}, \hat{\boldsymbol{\beta}}_{(l-1)}$ is $Gamma(y_i + \hat{\alpha}_{(l-1)}, \frac{1}{\hat{\alpha}_{(l-1)} + exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{(l-1)})})$. Therefore

$$E(ln\, u_i|y_i, \hat{\alpha}_{(l-1)}, \hat{\boldsymbol{\beta}}_{(l-1)}) = \psi(y_i + \hat{\alpha}_{(l-1)}) - ln(exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{(l-1)}) + \hat{\alpha}_{(l-1)})$$

and

$$E(u_i|y_i, \hat{\alpha}_{(l-1)}, \hat{\boldsymbol{\beta}}_{(l-1)}) = \frac{y_i + \hat{\alpha}_{(l-1)}}{\hat{\alpha}_{(l-1)} + exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{(l-1)})},$$

where $\psi(\cdot)$ denotes a digamma function (the first derivative of the log-gamma function).

At the l^{th} M-step the expected log-likelihood is maximized with respect to α and $\boldsymbol{\beta}$. The two parameters are present in different terms of the log-likelihood and hence two separate numerical maximizations are required. Note that if we adopted a direct approach the parameters would need to be maximized together which could lead to more numerical problems.

When there are random effects the model can be specified as follows:

$$y_{ij}|\mu_{ij}, u_{ij} \sim \; indep.\; Poisson(\mu_{ij} u_{ij})$$

$$u_{ij} \sim \; i.i.d.\; \Gamma(\alpha, \frac{1}{\alpha})$$

$$ln(\mu_{ij}|\mathbf{b_i}) = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{b_i}$$

$$\mathbf{b_i} \sim \; i.i.d.\; N_q(\mathbf{0}, \boldsymbol{\Sigma})$$

and the random effects $\mathbf{b}_i$ and u_{ij} are independent.

The unknown parameters α, $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ can be estimates by a nested EM algorithm. Such an algorithm has been proposed by Booth and Hobert (personal communication) for the overdispersed Poisson GLMM, and more generally by van Dyk (1999) who motivated it from the point of view of computational efficiency. The complete data for the outer loop consists of $\mathbf{y}$ and $\mathbf{b}$ and the outer EM algorithm is performed as outlined in Section 3.2.2 but with an EM maximization procedure for α and $\boldsymbol{\beta}$ as introduced above. The complete data log-likelihood has the form

$$ln\ L_u = c + \sum_{i,j}[ln\Gamma(y_{ij}+\alpha) - ln\Gamma(\alpha) + \alpha ln\alpha - (\alpha+y_{ij})ln(\alpha+\mu_{ij}) + y_{ij}ln(\mu_{ij})] +$$

$$\sum_i[-\frac{1}{2}ln|\boldsymbol{\Sigma}| - \frac{1}{2}\mathbf{b}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{b}_i]$$

and at the r^{th} E-step its conditional expectation is approximated by the Monte Carlo sum

$$\frac{1}{m}\sum_{k=1}^{m}[c + \sum_{i,j}[ln\Gamma(y_{ij}+\alpha) - ln\Gamma(\alpha) + \alpha ln\alpha - (\alpha+y_{ij})ln(\alpha+\mu_{ij}^{(k)}) + y_{ij}ln(\mu_{ij}^{(k)})] +$$

$$\frac{1}{m}\sum_{k=1}^{m}\sum_i[-\frac{1}{2}ln|\boldsymbol{\Sigma}| - \frac{1}{2}\mathbf{b}_i^{(k)T}\boldsymbol{\Sigma}^{-1}\mathbf{b}_i^{(k)}],$$

where $\mathbf{b}_i^{(k)}$ are generated from the conditional distribution of $\mathbf{b}_i|\mathbf{y}_i;\hat{\boldsymbol{\psi}}^{(r-1)}$ and $\mu_{ij}^{(k)} = exp(\mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{b}_i^{(k)})$. The first part of that sum is what needs to be maximized to obtain estimates of α and $\boldsymbol{\beta}$. Notice that each term in the first Monte Carlo sum is a log-likelihood for a negative binomial model with a different (but fixed) mean $\mu_{ij}^{(k)}$ and hence it can be subjected to an EM algorithm by augmenting the observed data $\mathbf{y}$ by $\mathbf{u}$.

In summary, the nested EM algorithm is as follows:

1. Select an initial estimate $\hat{\boldsymbol{\psi}}^{(0)}$. Set $r = 0$.

2. Increase r by 1.

E-step: For each subject i, $i = 1, ...n$ generate m random samples from the distribution of $\mathbf{b}_i|\mathbf{y}_i; \hat{\boldsymbol{\psi}}^{(r-1)}$ using rejection sampling and approximate $E(ln\ L_u(\mathbf{y}, \mathbf{b}; \boldsymbol{\psi})|\mathbf{y}; \hat{\boldsymbol{\psi}}^{(r-1)})$ by a Monte Carlo sum.

3. **M-step**:

3.1. Maximize with respect to the elements of $\boldsymbol{\Sigma}$ as usual.

3.2. Set $l = 0$, $\hat{\alpha}_{(0)} = \hat{\alpha}^{(r-1)}$ and $\hat{\boldsymbol{\beta}}_{(0)} = \hat{\boldsymbol{\beta}}^{(r-1)}$.

3.3. Increase l by 1.

Inner E-step: For any given $\mathbf{b}^{(k)}$ compute $E(ln\ L_u(\mathbf{y}, \mathbf{u}; \alpha, \boldsymbol{\beta})|\mathbf{y}; \hat{\alpha}_{(l-1)}, ; \hat{\boldsymbol{\beta}}_{(l-1)})$

3.4. **Inner M-step**: Find $\hat{\alpha}_{(l)}$ and $\hat{\boldsymbol{\beta}}_{(l)}$ to maximize the Monte Carlo sum of conditional expectations.

3.5. Iterate between (3.3) and (3.4) until convergence for α and $\boldsymbol{\beta}$ is achieved.

4. Iterate between (2) and (3) until convergence for $\boldsymbol{\psi}$ is achieved.

In the multivariate GLMM there are two or more response variables but the nested EM algorithm works in essentially the same way because of the conditional independence between the variables. In the next section we describe a bivariate GLMM for the pony data.

Analysis of Pony Data

The model is defined as follows:

y_{i1j} - j^{th} duration measurement on the i^{th} pony.

y_{i2j} - j^{th} spike burst rate measurement of the i^{th} pony.

$y_{i1j}|b_{i1} \sim$ *indep. Gamma* with mean μ_{i1j} and variance $\frac{\mu_{i1j}^2}{\nu}$.

$y_{i2j}|b_{i2}, u_{ij} \sim indep.\, Poisson(\mu_{i2j}u_{ij})$,

$u_{ij} \sim\ i.i.d.\ Gamma(\alpha, \frac{1}{\alpha})$,

$ln(\mu_{i1j}) = \mathbf{x}_{i1j}^T\boldsymbol{\beta}_1 + b_{i1}$,

$ln(\mu_{i2j}) = \mathbf{x}_{i2j}^T\boldsymbol{\beta}_2 + b_{i2}$,

$\mathbf{b}_i = (b_{i1}, b_{i2})' \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$, $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$.

Let d_{ijk} be 1 if drug k is administered to pony i at occasion j, $k = 1, ...6$, and 0 otherwise. Placebo is coded as drug 7 and will be the reference group. Let t denote time. Then the linear predictors are

$$\mathbf{x}_{i1j} = \beta_{10} + \beta_{11}d_{ij1} + \beta_{12}d_{ij2} + \beta_{13}d_{ij3} + \beta_{14}d_{ij4} + \beta_{15}d_{ij5} + \beta_{16}d_{ij6}$$

$$\mathbf{x}_{i2j} = \beta_{20} + \beta_{21}d_{ij1} + \beta_{22}d_{ij2} + \beta_{23}d_{ij3} + \beta_{24}d_{ij4} + \beta_{25}d_{ij5} + \beta_{26}d_{ij6}$$

$$+\beta_{27}t + \beta_{28}d_{ij1}t + \beta_{29}d_{ij2}t + \beta_{2,10}d_{ij3}t + \beta_{2,11}d_{ij4}t + \beta_{2,12}d_{ij5}t + \beta_{2,13}d_{ij6}t.$$

In the univariate analysis we initially considered the same linear predictor for both variables but the *time* $*$ *drug* interaction and the *time* main effect were clearly not significant for the duration response and were dropped from the model. Also, we performed fixed effects analysis using PROC GENMOD and fitted two random effects models on log-transformed variables using PROC MIXED for the complete data (all time points and all electrode groups). There was evidence of a three-way interaction between electrode group, time and drug. Describing this interaction requires estimating 76 regression parameters for a linear time trend as compared to 21 in the above specification. Recall that in the other data example we estimated up to five regression parameters, so the numerical maximization procedures in this example were expected to be much more complicated. The complete pony data set was also about ten times larger than the ethylene glycol data set and the time trends appeared complicated and not well described by simple linear or quadratic time effects. One might need to use splines to adequately describe the trends. Hence, we concentrated on a particular

research question, namely to describe differences in the immediate effects (up to 60 minutes after administration) of the drugs in the cecal base (corresponding to one of the electrode groups). In the analyses performed by Lester et al. (1998a, 1998b, 1999c) there was evidence that some drugs led to immediate significant increase in spike burst count or spike burst duration, while others took longer or did not lead to increase at all, and we decided to address that issue in our analysis.

We programmed the nested Monte Carlo EM and the adaptive Gaussian quadrature algorithms both for the joint and for the separate fits. We used a negative binomial distribution for the count response, a gamma distribution for the duration response and log link functions for both. (Although the log function is not the canonical link function for the Gamma response, it is convenient because it transforms the mean to the whole real line). As initial estimates we used the final estimates from the pseudo-likelihood approach using the $\%GLIMMIX$ macro but rounded only up to one significant digit after the decimal point. The macro does not allow specification of a negative binomial distribution and hence we fitted a Poisson distribution with an extra dispersion parameter.

The results from the analysis are summarized in Table 3.15. $MaxBFGS$ has convergence problems when the responses are fitted together and adaptive Gaussian quadrature is used so no results are presented for this case. We report the results using only one quadrature point because the estimates and their standard errors are essentially the same if the number of quadrature points is increased. Adaptive Gaussian quadrature with one quadrature point is equivalent to Laplacian approximation (Liu and Pierce, 1994) which is indicative that the analytical approximations work well for this data set. The simulation sample size for the separate fit of the Gamma response was increased at each iteration. Convergence was achieved when the simulation sample size was 31174 after 20 iterations and took about 1 hour. For the negative binomial response the final simulation sample size was still 100 after 542 iterations

and it took also about 1 hour to converge. The joint fit took about 21 hours until convergence with a final simulation sample size of 23381 and 201 iterations.

The results from the joint and from the separate fits are almost identical and the correlation between the two response variables is not significant ($\hat{\rho} = 0.71$, $SE = 0.50$). Notice that $\hat{\rho}$ is quite large and the fact that it is not significant may be partially due to the fact that there are only six subjects in the data set. Notice also that the standard error estimates from the three methods are similar with only some of the adaptive Gaussian quadrature standard errors being smaller than their Monte Carlo counterparts. The Monte Carlo standard error estimates did not show much variability as in the ethylene glycol example.

For both responses the coefficients for drug 1 and for drug 4 are significantly different from zero. This means that both drugs have significantly different immediate effects on the response variables than saline solution. Drug 1 leads to a significant decrease in the individual level of duration for one hour after the drug is administered and drug 4 is associated with a significant increase. Of all the interactions between drug and time only the one involving drug 1 for the count response is significant.

3.5 Additional Methods

The empirical Bayes posterior mode procedure, discussed in Fahrmeir and Tutz (1994, pp.233-238), can also be used to estimate the parameters. The Newton-Raphson equations for the extended model when a flat (or vague) prior is used and the dispersion parameters are set to 1.0, are essentially no more complicated than those for the generalized linear mixed models. Fahrmeir and Tutz, however, do not discuss the estimation of ϕ_1 and ϕ_2 when they are unknown. The dispersion parameters can be estimated together with $\boldsymbol{\Sigma}$ via maximum likelihood, treating the current estimates of the fixed and the random effects as the true values. Another possibility is to put noninformative priors on ϕ_1 and ϕ_2 and estimate them together with $\boldsymbol{\beta}$ and $\mathbf{b}$.

Table 3.15. Final maximum likelihood estimates for pony data

	Monte Carlo EM				Adaptive Gauss-Hermite quadrature	
	Joint fit		Separate fits		Separate fits	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
β_{10}	0.11	0.06	0.13	0.05	0.12	0.03
β_{11}	-0.23	0.05	-0.23	0.05	-0.23	0.05
β_{12}	0.11	0.05	0.11	0.05	0.11	0.05
β_{13}	0.01	0.05	0.02	0.05	0.02	0.05
β_{14}	0.32	0.05	0.32	0.05	0.32	0.05
β_{15}	0.10	0.05	0.10	0.05	0.10	0.05
β_{16}	0.18	0.05	0.18	0.05	0.18	0.05
β_{20}	4.25	0.22	4.30	0.22	4.31	0.19
β_{21}	-1.30	0.28	-1.31	0.28	-1.31	0.28
β_{22}	0.24	0.29	0.24	0.29	0.24	0.29
β_{23}	0.28	0.28	0.27	0.28	0.27	0.28
β_{24}	0.90	0.28	0.89	0.28	0.89	0.28
β_{25}	0.15	0.29	0.14	0.29	0.14	0.28
β_{26}	0.32	0.28	0.31	0.28	0.31	0.28
β_{27}	0.04	0.07	0.04	0.07	0.04	0.07
β_{28}	0.28	0.10	0.28	0.10	0.28	0.10
β_{29}	-0.05	0.11	-0.05	0.11	-0.05	0.10
$\beta_{2,10}$	-0.01	0.10	-0.01	0.10	-0.01	0.10
$\beta_{2,11}$	-0.17	0.10	-0.17	0.10	-0.17	0.10
$\beta_{2,12}$	-0.07	0.10	-0.06	0.10	-0.06	0.10
$\beta_{2,13}$	-0.14	0.10	-0.13	0.10	-0.13	0.10
ν	17.3	1.43	17.3	1.43	17.6	1.44
α	3.41	0.34	3.41	0.34	3.47	0.29
σ_1	0.10	0.03	0.10	0.03	0.09	0.03
σ_2	0.26	0.09	0.26	0.08	0.24	0.07
ρ	0.71	0.50	–	–	–	–

What noninformative priors should be used in order to avoid dealing with improper posteriors is a topic for further research. The question about the propriety of the posterior also arises when Gibbs sampling is applied for posterior mean estimation, which is yet another method that can be used for fitting the extended model.

CHAPTER 4
INFERENCE IN THE MULTIVARIATE GENERALIZED LINEAR MIXED MODEL

The estimates considered in the previous chapter are approximate maximum likelihood and hence confidence intervals and hypothesis tests can be constructed according to asymptotic maximum likelihood theory. Rigorous proof of the properties of the estimates requires check of the regularity conditions for consistency and asymptotic normality in the case of independent but not necessarily identically distributed random vectors. Such conditions have been established by Hoadley (1977) but are difficult to check for the generalized linear mixed model and its multivariate extension because of the lack of closed-form expression for the marginal likelihood. To our knowledge these conditions have not been verified for the generalized linear mixed model and we could not do that for the multivariate GLMM. Instead we assume that the regularity conditions are satisfied and rely on the general results for maximum likelihood estimates. Caution is applied to testing for the significance of variance components because when a parameter falls on the boundary of the parameter space the asymptotic distribution of its maximum likelihood estimate is no longer normal (Moran, 1971; Chant, 1974; Self and Liang, 1987). Score tests may be a good alternative to the usual Wald and likelihood-ratio tests because their asymptotic properties are retained on the boundary (Chant, 1974). Score test statistics are also computed under the null hypothesis and do not require fitting of more complicated models and hence can be used to check the conditional independence assumption.

Since there are no closed form expressions for the marginal likelihood, the score and the information matrix, numerical, stochastic or analytical approximations must

be used when computing test statistics and constructing confidence intervals. This aggravates the problem of determining actual error rates and coverage probabilities and requires the use of simulations to study the behaviour of the tests. Analytical and stochastic approximations can be improved by increasing the number of quadrature points or simulated sample sizes, but the precision of analytical approximations can not be directly controlled. For example, in the case of binary data pseudo-likelihood works well if the binomial denominator is large (Breslow and Clayton, 1993). But the latter depends only on the data, so for a particular data set, the approximation is either good or bad. In general the estimates based on analytical approximations are asymptotically biased.

In this chapter we concentrate on analytical and stochastic approximations of Wald, score and likelihood ratio statistics and study their performance for checking the conditional independence assumption. We also show how these approximations can be constructed for testing fixed effects and for estimating random effects, and consider them for checking the significance of variance components. Since the asymptotic maximum likelihood results hold when the number of subjects increases to infinity we focus on the ethylene glycol example. The pony data has only six subjects and as suggested in Chapter 3 inference concerning correlation between the two response variables is suspect.

Section 4.1 discusses Wald and likelihood ratio tests for testing the fixed effects. We briefly consider estimation of random effects and prediction of future observations in Section 4.2 The score approach is introduced in Section 4.3.1 by providing a historical overview. We then propose score tests for checking the conditional independence assumption (Section 4.3.2) and for testing the variance components (Section 4.3.2). The ethylene glycol example is used for illustration in Section 4.4 and Section 4.5

contains the results from a small simulation study for the performance of the proposed conditional independence test. The chapter concludes with discussion of future research topics.

4.1 Inference about Regression Parameters

The asymptotic properties of maximum likelihood estimates have been studied under a variety of conditions. The usual assumption is that the observations on which the maximum likelihood estimates are based are independent and identically distributed (Foutz, 1977), but results are available also for models in which the observations are independent but not identically distributed (Hoadley, 1971). In general, let $\mathbf{y}_1$, $\mathbf{y}_2$, ... $\mathbf{y}_n$ be independent random vectors with density or mass functions $f_1(\mathbf{y}_1, \boldsymbol{\psi})$, $f_2(\mathbf{y}_2, \boldsymbol{\psi})$, ... $f_n(\mathbf{y}_n, \boldsymbol{\psi})$ depending on a common unknown parameter vector $\boldsymbol{\psi}$. Then as $n \to \infty$ under certain regularity conditions the maximum likelihood estimator $\hat{\boldsymbol{\psi}}$ is consistent and asymptotically normal

$$\hat{\boldsymbol{\psi}} \xrightarrow{P} \boldsymbol{\psi}$$

$$n^{\frac{1}{2}}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \xrightarrow{d} N(\mathbf{0}, I^{-1}(\boldsymbol{\psi})),$$

where $I(\boldsymbol{\psi}) = lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E[-\frac{\partial^2 l_i(\mathbf{y}_i, \boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T}]$, $l_i(\mathbf{y}_i, \boldsymbol{\psi}) = ln f_i(\mathbf{y}_i, \boldsymbol{\psi})$.

Two basic principles of testing are directly based on the asymptotic distribution of the maximum likelihood estimate: the Wald test (Wald, 1943) and the likelihood ratio test (Neyman and Pearson, 1928). Let $\boldsymbol{\psi} = (\boldsymbol{\psi}_1^T, \boldsymbol{\psi}_2^T)^T$. The significance of a subset $\boldsymbol{\psi}_1$ of the parameter vector $\boldsymbol{\psi}$ can be tested by either one of them as follows. The null hypothesis is

$$H_0 : \boldsymbol{\psi}_1 = \mathbf{0} \tag{4.1}$$

where $\mathbf{0}$ is in the interior of the parameter space (In general the null hypothesis is $H_0 : \boldsymbol{\psi}_1 = \boldsymbol{\psi}_{10}$ but herein we consider the simpler test.) Let

$$I(\boldsymbol{\psi}) = \begin{pmatrix} I_{\boldsymbol{\psi}_1\boldsymbol{\psi}_1} & I_{\boldsymbol{\psi}_1\boldsymbol{\psi}_2} \\ I_{\boldsymbol{\psi}_2\boldsymbol{\psi}_1} & I_{\boldsymbol{\psi}_2\boldsymbol{\psi}_2} \end{pmatrix}.$$

Similarly

$$I^{-1}(\boldsymbol{\psi}) = \begin{pmatrix} I^{\boldsymbol{\psi}_1\boldsymbol{\psi}_1} & I^{\boldsymbol{\psi}_1\boldsymbol{\psi}_2} \\ I^{\boldsymbol{\psi}_2\boldsymbol{\psi}_1} & I^{\boldsymbol{\psi}_2\boldsymbol{\psi}_2} \end{pmatrix}.$$

The Wald test statistic is then

$$T_w = \hat{\boldsymbol{\psi}}_1^{\ T}(nI^{\boldsymbol{\psi}_1\boldsymbol{\psi}_1})\hat{\boldsymbol{\psi}}_1$$

with $\boldsymbol{\psi}_2$ in $I^{\boldsymbol{\psi}_1\boldsymbol{\psi}_1}$ replaced by the consistent maximum likelihood estimator $\tilde{\boldsymbol{\psi}}_2$ under the null hypothesis. Under H_0, $T_w \sim \chi^2_d$, where d is the dimension of the parameter vector $\boldsymbol{\psi}_1$. Because in the case of independent data $nI(\boldsymbol{\psi})$ may not be available and in random effects models the expected information matrix may be hard to calculate we use $\mathbf{J}(\boldsymbol{\psi}) = \sum_{i=1}^{n}(-\frac{\partial^2 l_i(y_i;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T})$ instead of $nI(\boldsymbol{\psi})$. Some authors argue that it is more appropriate to use the observed rather than the expected information (Efron and Hinkley, 1978) but unlike the expected information matrix, the observed information matrix is not guaranteed to be positive-definite. The latter problem is exacerbated when the numerical or stochastic approximations discussed in Section 3.2 are applied to approximate the observed information matrix $\mathbf{J}(\boldsymbol{\psi})$, which is not available in closed form. We already used Wald tests in Chapter 3 to test the significance of individual regression coeffients but we can also use Wald tests to check several regression coefficients simultaneously.

The likelihood ratio statistic for the hypothesis (4.1) is defined as follows. Let M_1 be the model with unknown parameter vector $\boldsymbol{\psi}$ and let M_2 be a reduced model with $\boldsymbol{\psi}_1 = 0$ and $\boldsymbol{\psi}_2$ held unrestricted. Let also l_1 and l_2 denote the maximized

log-likelihoods for models M_1 and M_2 respectively. Under H_0 as $n \to \infty$

$$T_{LR} = -2(l_2 - l_1) \sim \chi^2_d,$$

so the likelihood ratio and the Wald statistic have the same asymptotic distribution under the null hypothesis. The absence of closed-form expressions for l_1 and l_2 necessitates the use of either Gaussian quadrature or Monte Carlo approximations. Gaussian quadrature approximations are applied exactly as described in Section 3.2.1. To obtain Monte Carlo approximations m samples from the random effects distribution $\mathbf{b}_i \sim N_q(\mathbf{0}, \hat{\mathbf{\Sigma}})$ are generated. Then

$$l_1 \approx \frac{1}{m} \sum_{k=1}^{m} f(\mathbf{y_i}|\mathbf{b}_i^{(k)}; \hat{\boldsymbol{\beta}}, \hat{\phi}).$$

Here $\hat{\mathbf{\Sigma}}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$ are the final maximum likelihood estimates from the full model M_1. The same type of approximation is used for l_2 but evaluated at $\boldsymbol{\psi}_1 = 0$ and at the maximum likelihood estimator $\tilde{\boldsymbol{\psi}}$ under M_2.

Under ideal conditions for large enough number of quadrature points in the Gaussian quadrature algorithm and for large enough simulation sample size in the MCEM algorithm the approximate maximum likelihood estimates will be arbitrarily close to the true estimates. Also the additional approximations needed to compute the Wald and the likelihood ratio statistics above can be made very precise, hence the statistics should perform well for a large number of subjects. But in reality there can be problems because it is not clear how many quadrature points and what number of simulation samples are needed for adequate approximations.

Even if all approximations are adequate and the sample size is large enough the Wald and likelihood ratio tests can still run into problems if a parameter is on the boundary of the parameter space. This happens in testing the significance of a variance term. Hence Wald and likelihood should be used with caution in that situation. It has been proven by several authors (Moran, 1971; Chant, 1974, Self and Liang,

1987) that when a parameter is on the boundary of the parameter space the asymptotic distribution of the maximum likelihood estimator is no longer normal but rather a mixture of distributions. But the score tests as discussed in Section 4.3 are not affected and therefore may be a nice substitute for Wald and likelihood ratio tests.

4.2 Estimation of Random Effects

In some applications it is of interest to obtain estimates for the unobserved random effects. A natural point estimator is the conditional mean

$$E[\mathbf{b}_i|\mathbf{y}_i, \hat{\boldsymbol{\psi}}] = \frac{\int \mathbf{b}_i f(\mathbf{y}_i|\mathbf{b}_i, \hat{\boldsymbol{\psi}}) f(\mathbf{b}_i; \hat{\boldsymbol{\psi}}) d\mathbf{b}_i}{\int f(\mathbf{y}_i|\mathbf{b}_i, \hat{\boldsymbol{\psi}}) f(\mathbf{b}_i; \hat{\boldsymbol{\psi}}) d\mathbf{b}_i}$$

which is not available in closed form but can be approximated either numerically or stochastically. Gauss-Hermite quadrature involves two separate approximations of the numerator and the denominator, while the stochastic approximation is via the simple Monte Carlo sum

$$\frac{1}{m}\sum_{k=1}^{m} \mathbf{b}_i^{(k)},$$

where $\mathbf{b}_i^{(1)}$, ... $\mathbf{b}_i^{(m)}$ are simulated from the conditional distribution of $\{\mathbf{b}_i|\mathbf{y}_i, \hat{\boldsymbol{\psi}}\}$ using a technique such as rejection sampling. Note that this approximation is performed anyway in the MCEM algorithm and hence the random effects estimate is obtained at no extra cost. Estimates of the random effects are needed for prediction. For example one might be interested in obtaining an estimate for the linear predictor for particular subject:

$$\hat{\boldsymbol{\eta}}_i = \mathbf{X}_i^T \hat{\boldsymbol{\beta}} + \mathbf{Z}_i^T \hat{\mathbf{b}}_i.$$

The question with obtaining the variance of the random effects estimate is not straightforward. The simplest approach is to use the variance

$$Var(\mathbf{b}_i|\mathbf{y}_i, \boldsymbol{\psi})$$

evaluated at $\hat{\psi}$, but this 'naive' estimate may underestimate the true variance as it does not account for the sampling variability of $\hat{\psi}$. As an alternative Booth and Hobert (1998) suggested using a 'conditional mean square error of prediction' as a measure of prediction variance. Their approach can also be applied to the multivariate GLMM.

Note that in the two 'real-life' examples that we consider estimation of the random effects is not of particular interest. The subjects are mice and ponies respectively and their individual characteristics are a source of variability that needs to be accounted for but not necessarily precisely estimated for each subject. In other applications, for example in small area estimation, prediction of the random effects is a very important objective of the analysis.

4.3 Inference Based on Score Tests

4.3.1 General Theory

An overview of the historical development of the score test is provided in a research paper by Bera and Bilias (1999). Rao (1947) was the first to introduce the fundamental principle of testing based on the score function as an alternative to likelihood ratio and Wald tests. Let $\mathbf{y}_1$, $\mathbf{y}_2$, ... $\mathbf{y}_n$ be i.i.d. observations with density $f(\mathbf{y}_i, \boldsymbol{\psi})$. Denote the joint log-likelihood, score and expected information matrix of those observations by $l(\boldsymbol{\psi})$, $\mathbf{s}(\boldsymbol{\psi})$ and $\mathbf{I}(\boldsymbol{\psi})$ respectively. Suppose that the interest is in testing a simple hypothesis against a local alternative

$$H_0 : \boldsymbol{\psi} = \boldsymbol{\psi}_0 \; vs \; H_a : \boldsymbol{\psi} = \boldsymbol{\psi}_0 + \boldsymbol{\Delta},$$

where $\boldsymbol{\psi} = (\psi_1, ...\psi_p)^T$ and $\boldsymbol{\Delta} = (\Delta_1, ...\Delta_p)^T$. Then the score test is based on

$$T_s = \mathbf{s}(\boldsymbol{\psi}_0)^T \mathbf{I}(\boldsymbol{\psi}_0)^{-1} \mathbf{s}(\boldsymbol{\psi}_0)$$

which has an asymptotic χ^2_p distribution. If the null hypothesis is composite, that is

$$H_0 : \mathbf{h}(\boldsymbol{\psi}) = \mathbf{c},$$

where $\mathbf{h}(\boldsymbol{\psi})$ is a $r \times 1$ vector function of $\boldsymbol{\psi}$ with $r \leq p$ restrictions, and $\mathbf{c}$ is a known constant vector, Rao (1947) suggested using

$$T_s(\tilde{\boldsymbol{\psi}}) = \mathbf{s}(\tilde{\boldsymbol{\psi}})^T \mathbf{I}(\tilde{\boldsymbol{\psi}})^{-1} \mathbf{s}(\tilde{\boldsymbol{\psi}}),$$

where $\tilde{\boldsymbol{\psi}}$ is the restricted maximum likelihood estimate of $\boldsymbol{\psi}$, that is, the estimate obtained by maximizing the log-likelihood function $l(\boldsymbol{\psi})$ under H_0. $T_s(\tilde{\boldsymbol{\psi}})$ has an asymptotic χ^2_r distribution.

Independently of Rao's work Neyman (1959) proposed $C(\alpha)$ tests. These tests are specifically designed to deal with hypothesis testing of a parameter of primary interest in the presence of nuisance parameters and are more general than Rao's score tests in that any $\sqrt{n}$-consistent estimates for the nuisance parameters can be used, not only the maximum likelihood estimators. By design $C(\alpha)$ tests maximize the slope of the limiting power function under local alternatives to the null hypothesis.

Neyman assumed the same setup as in Rao's score test but considered the simple null hypothesis

$$H_0 : \psi_1 = \psi_{10},$$

where $\boldsymbol{\psi} = (\psi_1, \boldsymbol{\psi}_2^T)^T$. Notice that ψ_1 is a scalar. The score vector and the information matrix are partitioned as follows

$$\mathbf{s}(\boldsymbol{\psi}) = \begin{pmatrix} s_{\psi_1}(\psi_1, \boldsymbol{\psi}_2) \\ \mathbf{s}_{\psi_2}(\psi_1, \boldsymbol{\psi}_2) \end{pmatrix}$$

$$\mathbf{I}(\boldsymbol{\psi}) = \begin{pmatrix} I_{\psi_1\psi_1}(\psi_1, \boldsymbol{\psi}_2) & \mathbf{I}_{\psi_1 \boldsymbol{\psi}_2}(\psi_1, \boldsymbol{\psi}_2) \\ \mathbf{I}_{\boldsymbol{\psi}_2 \psi_1}(\psi_1, \boldsymbol{\psi}_2) & \mathbf{I}_{\boldsymbol{\psi}_2 \boldsymbol{\psi}_2}(\psi_1, \boldsymbol{\psi}_2) \end{pmatrix}.$$

Then the $C(\alpha)$ score statistic is

$$C(\alpha) = [s_{\psi_1}(\psi_{10}, \tilde{\boldsymbol{\psi}}_2) - \mathbf{I}_{\psi_1 \boldsymbol{\psi}_2}(\psi_{10}, \tilde{\boldsymbol{\psi}}_2) \mathbf{I}^{-1}_{\boldsymbol{\psi}_2 \boldsymbol{\psi}_2}(\tilde{\boldsymbol{\psi}}_2, \tilde{\boldsymbol{\psi}}_2) \mathbf{s}_{\boldsymbol{\psi}_2}(\psi_{10}, \tilde{\boldsymbol{\psi}}_2)]^{\mathbf{T}}$$

$$\times[I_{\psi_1\psi_1}(\psi_{10},\tilde{\boldsymbol{\psi}}_2)-\mathbf{I}_{\psi_1\boldsymbol{\psi}_2}(\psi_{10},\tilde{\boldsymbol{\psi}}_2)\mathbf{I}^{-1}_{\boldsymbol{\psi}_2\boldsymbol{\psi}_2}(\tilde{\boldsymbol{\psi}}_2,\tilde{\boldsymbol{\psi}}_2)\mathbf{I}_{\boldsymbol{\psi}_2\psi_1}(\psi_{10},\tilde{\boldsymbol{\psi}}_2)]^{-1}$$

$$\times[s_{\psi_1}(\psi_{10},\tilde{\boldsymbol{\psi}}_2)-\mathbf{I}_{\psi_1\boldsymbol{\psi}_2}(\psi_{10},\tilde{\boldsymbol{\psi}}_2)\mathbf{I}^{-1}_{\boldsymbol{\psi}_2\boldsymbol{\psi}_2}(\tilde{\boldsymbol{\psi}}_2,\tilde{\boldsymbol{\psi}}_2)\mathbf{s}_{\boldsymbol{\psi}_2}(\psi_{10},\tilde{\boldsymbol{\psi}}_2)],$$

where $\tilde{\boldsymbol{\psi}}_2$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\psi}_2$. The Neyman's $C(\alpha)$ statistic reduces to the Rao's score statistic when $\tilde{\boldsymbol{\psi}}_2$ is the maximum likelihood estimate of $\boldsymbol{\psi}_2$ under H_0.

Buhler and Puri (1966) extended the asymptotic and local optimality of Neyman's $C(\alpha)$ statistic for a vector valued $\boldsymbol{\psi}$ and in the case of independent but not necessarily i.i.d. random variables. They assumed that $\boldsymbol{\psi}_{10}$ was interior to an open set in the parameter space but as pointed out later by Moran (1971) and Chant (1974) this restriction is unnecessary.

Chant (1974) showed that when the parameter is on the boundary of a closed parameter space, the score test retains its asymptotic properties while the asymptotic distributional forms of the test statistics based on the maximum likelihood estimators are no longer χ^2. In addition to this advantage of the score test it has the computational advantage that only estimates under the null hypothesis are needed to compute the test statistic. We now use this feature to propose tests for checking the conditional independence assumption.

4.3.2 Testing the Conditional Independence Assumption

Difficulties in testing the conditional independence of the response variables arise because of the need to specify more complicated models for the joint response distribution. Even if score tests are used (which as discussed do not require fitting of more complicated models) extended models need to be specified and the form of the score function and of the information matrix need to be derived. We again consider the bivariate GLMM case for simplicity. A convenient way to introduce conditional

dependence is to use one of the response variables as a covariate in the linear predictor for the other response variable. In the bivariate case this leads to considering the following model

$$y_{i1j}|\mathbf{y}_{i2}, \mathbf{b}_{i1} \sim \ indep\ f_1(y_{i1j}|y_{i2j}, \mathbf{b}_{i1}; \boldsymbol{\beta}_1, \gamma, \phi_1)$$

$$y_{i2j}|\mathbf{b}_{i2} \sim \ indep\ f_2(y_{i2j}|\mathbf{b}_{i2}; \boldsymbol{\beta}_2, \phi_2)$$

$$g_1(\mu_{i1j}) = \mathbf{x}_{i1j}^T\boldsymbol{\beta}_1 + \gamma y_{i2j} + \mathbf{z}_{i1j}^T\mathbf{b}_{i1}$$

$$g_2(\mu_{i2j}) = \mathbf{x}_{i2j}^T\boldsymbol{\beta}_2 + \mathbf{z}_{i2j}^T\mathbf{b}_{i2}$$

$$\mathbf{b}_i = \begin{pmatrix} \mathbf{b}_{i1} \\ \mathbf{b}_{i2} \end{pmatrix} \sim i.i.d.\, \mathbf{MVN}(\mathbf{0}, \boldsymbol{\Sigma}) = \mathbf{MVN}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right).$$

In general this setup leads to a complicated form of conditional dependence which is hard to interpret if there is no natural ordering to the two responses. The case $\gamma = 0$ corresponds to conditional independence but testing $\gamma = 0$ in the above model is performed against a complicated alternative on the marginal scale. When the identity link function is used for the first response, conditional on the random effects $Cov(y_{i1j}, y_{i2j}) = \gamma Var(y_{i2j})$ and the test is a test of conditional uncorrelatedness of the two outcomes. If both outcomes are normally distributed then this is truly a test of conditional independence.

An interesting case to consider in view of the simulated data example and the ethylene glycol application is when one of the responses is normally distributed and the other one has a Bernoulli distribution. Let in the above general specification f_1 be the normal density function and f_2 be the Bernoulli probability function. Also assume that the random effects consist of two random intercepts and let

$$\mu_{i1j} = \mathbf{x}_{i1j}^T\boldsymbol{\beta}_1 + \gamma y_{i2j}^* + b_{i1},$$

where $y_{i2j}^* = 2y_{i2j} - 1$. Then

$$E(y_{i1j}|y_{i2j}^* = 1, \mathbf{b}_i) = \mathbf{x}_{i1j}^T\boldsymbol{\beta}_1 + b_{i1} + \gamma$$

$$E(y_{i1j}|y^*_{i2j} = -1, \mathbf{b}_i) = \mathbf{x}^T_{i1j}\boldsymbol{\beta}_1 + b_{i1} - \gamma$$

and hence testing $H_0 : \gamma = 0$ against $H_1 : \gamma \neq 0$ is equivalent to testing for location shift in the conditional distribution of the normal response.

The score test statistic is as follows:

$$T_s = \frac{s^2_\gamma}{I_{\gamma\gamma} - \mathbf{I}_{\gamma,-\gamma}\mathbf{I}^{-1}_{-\gamma,-\gamma}\mathbf{I}_{-\gamma,\gamma}},$$

where s_γ is the element of the score vector corresponding to γ and $\mathbf{I}$ is the expected information matrix. Note that even in this simple case neither the score nor the expected information matrix have closed form expressions and hence the score statistic must be approximated. We again consider Gaussian quadrature and Monte Carlo approximations.

The log-likelihood is $lnL = \sum_{i=1}^n lnL_i$, where

$$lnL_i = ln \int f_1(\mathbf{y}_{i1}|\mathbf{y}_{i2}, \mathbf{b}_{i1}; \boldsymbol{\beta}_1, \gamma, \phi_1) f_2(\mathbf{y}_{i2}|\mathbf{b}_{i2}; \boldsymbol{\beta}_2, \phi_2) f(\mathbf{b}_i; \boldsymbol{\Sigma}) d\mathbf{b}_i$$

$$f_1(\mathbf{y}_{i1}|\mathbf{y}_{i2}, \mathbf{b}_{i1}; \boldsymbol{\beta}_1, \gamma, \phi_1) = (2\pi\sigma^2)^{-\frac{n_i}{2}} exp\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(y_{i1j} - \mathbf{x}^T_{i1j}\boldsymbol{\beta}_1 - b_{i1} - \gamma y^*_{i2j})^2\}$$

$$f_2(\mathbf{y}_{i2}|\mathbf{b}_{i2}; \boldsymbol{\beta}_2, \phi_2) = \frac{exp(\sum_{j=1}^{n_i} y_{i2j}(\mathbf{x}^T_{i2j}\boldsymbol{\beta}_2 + b_{i2}))}{\prod_{j=1}^{n_i}(1 + exp(\mathbf{x}^T_{i2j}\boldsymbol{\beta}_2 + b_{i2}))}$$

$$f(\mathbf{b}_i; \boldsymbol{\Sigma}) = |2\pi\boldsymbol{\Sigma}|^{-\frac{1}{2}} exp\{-\frac{1}{2}\mathbf{b}^T_i\boldsymbol{\Sigma}^{-1}\mathbf{b}_i\}.$$

For the Monte Carlo approximation we notice that under the assumption of interchangeability of the integral and differential signs

$$s_\gamma = \sum_{i=1}^n \frac{\partial}{\partial\gamma} ln \int f_1(\mathbf{y}_{i1}|\mathbf{b}_{i1}, \mathbf{y}_{i2}, \boldsymbol{\beta}_1, \gamma, \phi_1) f_2(\mathbf{y}_{i2}|\mathbf{b}_{i2}, \boldsymbol{\beta}_2, \phi_2) f(\mathbf{b}_i, \boldsymbol{\Sigma}) d\mathbf{b}_i =$$

$$= \sum_{i=1}^n \frac{\int \frac{\partial}{\partial\gamma}(f_1(\mathbf{y}_{i1}|\mathbf{b}_{i1}, \mathbf{y}_{i2}, \boldsymbol{\beta}_1, \gamma, \phi_1) f_2(\mathbf{y}_{i2}|\mathbf{b}_{i2}, \boldsymbol{\beta}_2, \phi_2) f(\mathbf{b}_i, \boldsymbol{\Sigma})) d\mathbf{b}_i}{\int f_1(\mathbf{y}_{i1}|\mathbf{b}_{i1}, \mathbf{y}_{i2}, \boldsymbol{\beta}_1, \gamma, \phi_1) f_2(\mathbf{y}_{i2}|\mathbf{b}_{i2}, \boldsymbol{\beta}_2, \phi_2) f(\mathbf{b}_i, \boldsymbol{\Sigma}) d\mathbf{b}_i} =$$

$$= \sum_{i=1}^{n} \frac{\int \frac{\partial}{\partial\gamma} ln f_1(\mathbf{y}_{i1}|\mathbf{b}_{i1}, \mathbf{y}_{i2}, \boldsymbol{\beta}_1, \gamma, \phi_1) f(\mathbf{y}_i, \mathbf{b}_i; \boldsymbol{\psi}) d\mathbf{b}_i}{f(\mathbf{y}_i; \boldsymbol{\psi})} =$$

$$\sum_{i=1}^{n} \int \frac{\partial}{\partial\gamma} ln f(\mathbf{y}_{i1}|\mathbf{b}_{i1}, \mathbf{y}_{i2}, \boldsymbol{\beta}_1, \gamma, \phi_1) f(\mathbf{b}_i|\mathbf{y}_i; \boldsymbol{\psi}) d\mathbf{b}_i =$$

$$\sum_{i=1}^{n} E(\frac{\partial}{\partial\gamma} ln f_1(\mathbf{y}_{i1}|\mathbf{y}_{i2}, \mathbf{b}_i)|\mathbf{y}_i).$$

The expectation above is taken with respect to the conditional distribution of the random effects given the response vector. Differentiating with respect to γ

$$\frac{\partial}{\partial\gamma} ln f_1(\mathbf{y}_{i1}|\mathbf{y}_{i2}, \mathbf{b}_i) = \frac{1}{\sigma^2} \sum_{j=1}^{n_i} y^*_{i2j}(y_{i1j} - \mathbf{x}^T_{i1j}\boldsymbol{\beta}_1 - b_{i1} - \gamma y^*_{i2j})$$

and therefore

$$E(\frac{\partial}{\partial\gamma} ln f_1(\mathbf{y}_{i1}|\mathbf{y}_{i2}, \mathbf{b}_i)) = \frac{1}{\sigma^2} \sum_{j=1}^{n_i} y^*_{i2j}(y_{i1j} - \mathbf{x}^T_{i1j}\boldsymbol{\beta}_1 - E(b_{i1}|\mathbf{y}_i) - \gamma y^*_{i2j}).$$

So, to approximate the score under the null hypothesis we only need to approximate the conditional mean $E(b_{i1}|\mathbf{y}_i)$ by the Monte Carlo sum $\frac{1}{m}\sum_{k=1}^{m} \mathbf{b}_i^{(k)}$, where $\mathbf{b}_i^{(k)}$, $k = 1, ...m$ are generated for the estimation of the standard errors in the MCEM algorithm (Section 3.2.2). The elements of the observed information matrix J, which can be used in place of I, can also be approximated using the Louis's method. $J_{-\gamma,-\gamma}$ is available from the MCEM algorithm and only $J_{\gamma,\gamma}$, $I_{\gamma,-\gamma}$ and $J_{-\gamma,\gamma}$ need to be computed. The latter can also be performed in the procedure for finding the standard errors of the estimates in the MCEM algorithm.

Gaussian quadrature using numerical derivatives involves approximating the log-likelihood once and then numerically differentiating with respect to γ and the other parameters to obtain the score and the observed information matrix. Denote the Gauss-Hermite quadrature approximation of the log-likelihood by l^{GQ} and let $\mathbf{s}_\gamma^{GQ} =$

$\frac{\partial l^{GQ}}{\partial \gamma}$ and $\mathbf{J}^{GQ}_{\gamma,-\gamma} = -\frac{\partial^2 l^{GQ}}{\partial \gamma \partial -\gamma}$. Then the approximation to the score statistic is

$$\frac{(s^{GQ}_{\gamma})^2}{J^{GQ}_{\gamma\gamma} - \mathbf{J}^{GQ}_{\gamma,-\gamma}\mathbf{J}^{GQ}_{-\gamma,-\gamma}{}^{-1}\mathbf{J}^{GQ}_{-\gamma,\gamma}}.$$

The performance of the score statistics for conditional independence is studied in more detail in Section 4.5. When there are more than two response variables this approach to testing for departure from conditional independence becomes very complicated and not easily interpretable. It is also not easy to decide which variable to use in the linear predictor for the other one, unless there is a natural ordering. This issue is discussed in more detail for the ethylene glycol example.

4.3.3 Testing the Significance of Variance Components

Global Variance Components Test

Lin (1997) proposed a global variance component test for testing the significance of all variance components in the univariate GLMM which can be extended for the multivariate GLMM. The null hypothesis for the global test is $H_0 : \boldsymbol{\delta} = \mathbf{0}$, where $\boldsymbol{\delta}$ is the vector of all variance components for the random effects. Suppose for simplicity that $\phi_1 = \phi_2 = 1$ and that there are two response variables. The generalizations to arbitrary ϕ_1 and ϕ_2, and to more than two response variables are straightforward. The form of the score test statistic is

$$T_s(\tilde{\boldsymbol{\beta}}) = \mathbf{s}_{\boldsymbol{\delta}}(\tilde{\boldsymbol{\beta}})^T(\mathbf{I}_{\boldsymbol{\delta}\boldsymbol{\delta}}(\tilde{\boldsymbol{\beta}}) - \mathbf{I}_{\boldsymbol{\delta}\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}})\mathbf{I}^{-1}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}})\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\delta}}(\tilde{\boldsymbol{\beta}}))^{-1}\mathbf{s}_{\boldsymbol{\delta}}(\tilde{\boldsymbol{\beta}}),$$

where $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_1^T)^T$ and $\tilde{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_2$ are the maximum likelihood estimates under H_0, i.e. the maximum likelihood estimators from the two separate fixed effects generalized linear models for the two response variables. Under H_0, $T_s(\tilde{\boldsymbol{\beta}})$ has an asymptotic χ^2_d distribution, where d is the number of variance-covariance parameters for the random effects.

In the univariate GLMM considered by Lin the r^{th} element of the score vector has the form

$$\mathbf{s}_{\delta_r}(\tilde{\boldsymbol{\beta}}) = \frac{1}{2}\sum_{i=1}^{n}\{(\mathbf{y}_i-\boldsymbol{\mu}_i)^T\Delta_i^{-1}\mathbf{W}_i\mathbf{Z}_i\dot{\boldsymbol{\Sigma}}^r\mathbf{Z}_i^T\mathbf{W}_i\Delta_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i) - tr(\mathbf{W}_{0i}\mathbf{Z}_i\dot{\boldsymbol{\Sigma}}^r\mathbf{Z}_i^T)\},$$

where $g(\boldsymbol{\mu}_i) = \mathbf{X}_i\boldsymbol{\beta}$, $\boldsymbol{\Sigma} = Var(\mathbf{b}_i)$ and $\dot{\boldsymbol{\Sigma}}^r = \frac{\partial\boldsymbol{\Sigma}}{\partial\delta_r}\mid_{\boldsymbol{\delta}=\mathbf{0}}$. The matrices $\boldsymbol{\Delta}_i$, $\mathbf{W}_i$ and $\mathbf{W}_{0i}$ are diagonal with elements $\Delta_{ij} = \frac{1}{g'(\mu_{ij})}$, $w_{ij} = (V(\mu_{ij})\{g'(\mu_{ij})\}^2)^{-1}$ and $w_{0ij} = w_{ij} + e_{ij}(y_{ij}-\mu_{ij})$ where $e_{ij} = \frac{V'(\mu_{ij})g'(\mu_{ij})+V(\mu_{ij})g''(\mu_{ij})}{V^2(\mu_{ij})[g'(\mu_{ij})]^3}$ in general and $e_{ij} = 0$ for canonical link functions. The subscript j refers to the j^{th} observation on the i^{th} subject.

Following step by step Lin's derivation for the univariate GLMM, the corresponding r^{th} element of the score function for the multivariate GLMM is $\mathbf{s}_{\boldsymbol{\delta}_r}(\tilde{\boldsymbol{\beta}}) =$

$$\frac{1}{2}\sum_{i=1}^{n}\{(\mathbf{y}_{i1}-\boldsymbol{\mu}_{i1})^T\Delta_{i1}^{-1}\mathbf{W}_{i1}\mathbf{Z}_{i1}\dot{\boldsymbol{\Sigma}}_{11}^r\mathbf{Z}_{i1}^T\mathbf{W}_{i1}\Delta_{i1}^{-1}(\mathbf{y}_{i1}-\boldsymbol{\mu}_{i1}) - tr(\mathbf{W}_{0i1}\mathbf{Z}_{i1}\dot{\boldsymbol{\Sigma}}_{11}^r\mathbf{Z}_{i1}^T)\}$$

$$+\frac{1}{2}\sum_{i=1}^{n}\{(\mathbf{y}_{i2}-\boldsymbol{\mu}_{i2})^T\Delta_{i2}^{-1}\mathbf{W}_{i2}\mathbf{Z}_{i2}\dot{\boldsymbol{\Sigma}}_{22}^r\mathbf{Z}_{i2}^T\mathbf{W}_{i2}\Delta_{i2}^{-1}(\mathbf{y}_{i2}-\boldsymbol{\mu}_{i2}) - tr(\mathbf{W}_{0i2}\mathbf{Z}_{i2}\dot{\boldsymbol{\Sigma}}_{22}^r\mathbf{Z}_{i2}^T)\}$$

$$+\sum_{i=1}^{n}\{(\mathbf{y}_{i2}-\boldsymbol{\mu}_{i2})^T\Delta_{i2}^{-1}\mathbf{W}_{i2}\mathbf{Z}_{i2}\dot{\boldsymbol{\Sigma}}_{12}^r\mathbf{Z}_{i1}^T\mathbf{W}_{i1}\Delta_{i1}^{-1}(\mathbf{y}_{i1}-\boldsymbol{\mu}_{i1})\}, \qquad (4.2)$$

where the subscripts 1 and 2 refer to the parts of the vectors and matrices corresponding to the first and to the second variable respectively.

The proof is as follows. Let

$$l_i(\mathbf{b}_i) = l_{i1}(\mathbf{b}_{i1}) + l_{i2}(\mathbf{b}_{i2}) = ln(f_1(\mathbf{y}_{i1}|\mathbf{b}_{i1};\boldsymbol{\beta}_1)f_2(\mathbf{y}_{i2}|\mathbf{b}_{i2};\boldsymbol{\beta}_2)).$$

Then the marginal likelihood for the i^{th} subject is

$$f_i(\mathbf{y}_i;\boldsymbol{\beta},\boldsymbol{\Sigma}) = E_{\mathbf{b}_i}[exp(l_i(\mathbf{b}_i))],$$

where the expectation is taken with respect to the marginal distribution of $\mathbf{b}_i$. Expanding the integrand in a multivariate Taylor series around $\mathbf{b}_i = \mathbf{0}$ we get

$$exp(l_i(\mathbf{b}_i)) = exp(l_i(\mathbf{0}))[1 + \frac{\partial l_i(\mathbf{0})}{\partial \mathbf{b}_i^T}\mathbf{b}_i + \frac{1}{2}\mathbf{b}_i^T(\frac{\partial l_i(\mathbf{0})}{\partial \mathbf{b}_i}\frac{\partial l_i(\mathbf{0})}{\partial \mathbf{b}_i^T} + \frac{\partial^2 l_i(\mathbf{0})}{\partial \mathbf{b}_i \partial \mathbf{b}_i^T})\mathbf{b}_i + \epsilon_i],$$

where ϵ_i contains third and higher order terms of $\mathbf{b}_i$. Notice that

$$\frac{\partial l_i(\mathbf{b}_i)}{\partial \mathbf{b}_i} = \mathbf{Z}_i^T \frac{\partial l_i(\mathbf{b}_i)}{\partial \boldsymbol{\eta}_i}$$

and

$$\frac{\partial^2 l_i(\mathbf{b}_i)}{\partial \mathbf{b}_i \partial \mathbf{b}_i^T} = \mathbf{Z}_i^T \frac{\partial^2 l_i(\mathbf{b}_i)}{\partial \boldsymbol{\eta}_i \partial \boldsymbol{\eta}_i^T}\mathbf{Z}_i,$$

where $\boldsymbol{\eta}_i$ is the vector of linear predictors for the i^{th} subject. Then taking expectation and using the moment assumptions for $\mathbf{b}_i$

$$E_{\mathbf{b}_i}[exp(l_i(\mathbf{b}_i))] = exp(l_i(\mathbf{0}))[1 + \frac{1}{2}tr(\mathbf{Z}_i^T(\frac{\partial l_i(\mathbf{0})}{\partial \boldsymbol{\eta}_i}\frac{\partial l_i(\mathbf{0})}{\partial \boldsymbol{\eta}_i^T} + \frac{\partial^2 l_i(\mathbf{0})}{\partial \boldsymbol{\eta}_i \partial \boldsymbol{\eta}_i^T})\mathbf{Z}_i\boldsymbol{\Sigma}) + r_i]$$

and the marginal log-likelihood for the ith subject is

$$ln f_i(\mathbf{y}_i; \boldsymbol{\beta}, \boldsymbol{\Sigma}) = l_i(\mathbf{0}) + \frac{1}{2}tr(\mathbf{Z}_i^T(\frac{\partial l_i(\mathbf{0})}{\partial \boldsymbol{\eta}_i}\frac{\partial l_i(\mathbf{0})}{\partial \boldsymbol{\eta}_i^T} + \frac{\partial^2 l_i(\mathbf{0})}{\partial \boldsymbol{\eta}_i \partial \boldsymbol{\eta}_i^T})\mathbf{Z}_i\boldsymbol{\Sigma}) + r_i.$$

Here r_i contains terms that are products of variance components and its derivative will be 0 when evaluated under H_0. Now we must take into consideration that there are two response variables. Because l_{i1} and l_{i2} depend on different sets of random effects

$$\frac{\partial l_i(\mathbf{b}_i)}{\partial \boldsymbol{\eta}_i} = \begin{pmatrix} \frac{\partial l_{i1}(\mathbf{b}_{i1})}{\partial \boldsymbol{\eta}_{i1}} \\ \\ \frac{\partial l_{i2}(\mathbf{b}_{i2})}{\partial \boldsymbol{\eta}_{i2}} \end{pmatrix}$$

and

$$\frac{\partial^2 l_i(\mathbf{b}_i)}{\partial \boldsymbol{\eta}_i \partial \boldsymbol{\eta}_i^T} = \begin{pmatrix} \frac{\partial^2 l_{i1}(\mathbf{b}_{i1})}{\partial \boldsymbol{\eta}_{i1} \partial \boldsymbol{\eta}_{i1}^T} & \mathbf{0} \\ \\ \mathbf{0} & \frac{\partial^2 l_{i2}(\mathbf{b}_{i2})}{\partial \boldsymbol{\eta}_{i2} \partial \boldsymbol{\eta}_{i2}^T} \end{pmatrix}.$$

Then

$$lnf_i(\mathbf{y}_i;\boldsymbol{\psi}) = l_{i1}(\mathbf{0}) + l_{i2}(\mathbf{0}) +$$

$$+\frac{1}{2}tr(\mathbf{Z}_{i1}^T(\frac{\partial l_{i1}(\mathbf{0})}{\partial \boldsymbol{\eta}_{i1}}\frac{\partial l_{i1}(\mathbf{0})}{\partial \boldsymbol{\eta}_{i1}^T} + \frac{\partial^2 l_{i1}(\mathbf{0})}{\partial \boldsymbol{\eta}_{i1}\partial \boldsymbol{\eta}_{i1}^T})\mathbf{Z}_{i1}\boldsymbol{\Sigma}_{11})$$

$$+\frac{1}{2}tr(\mathbf{Z}_{i2}^T(\frac{\partial l_{i2}(\mathbf{0})}{\partial \boldsymbol{\eta}_{i2}}\frac{\partial l_{i2}(\mathbf{0})}{\partial \boldsymbol{\eta}_{i2}^T} + \frac{\partial^2 l_{i2}(\mathbf{0})}{\partial \boldsymbol{\eta}_{i2}\partial \boldsymbol{\eta}_{i2}^T})\mathbf{Z}_{i2}\boldsymbol{\Sigma}_{22})$$

$$+tr(\mathbf{Z}_{i2}^T\frac{\partial l_{i2}(\mathbf{0})}{\partial \boldsymbol{\eta}_{i2}}\frac{\partial l_{i1}(\mathbf{0})}{\partial \boldsymbol{\eta}_{i1}^T}\mathbf{Z}_{i1}\boldsymbol{\Sigma}_{12}).$$

To obtain (4.2) one uses the fact that l_{i1} and l_{i2} do not depend on $\boldsymbol{\delta}$ and that for exponential family distributions

$$\frac{\partial l_{ik}(\mathbf{0})}{\partial \boldsymbol{\eta}_{ik}} = \mathbf{W}_{ik}\Delta_{ik}^{-1}(\mathbf{y}_{ik} - \boldsymbol{\mu}_{ik})$$

$$\frac{\partial^2 l_{ik}(\mathbf{0})}{\partial \boldsymbol{\eta}_{ik}\partial \boldsymbol{\eta}_{ik}^T} = \mathbf{W}_{0ik}$$

for $k = 1, 2$.

Notice that in the bivariate GLMM it is likely that the two responses require different variance components, in which case the expressions for the elements of the score vector above simplify. Suppose that the random effects variance-covariance matrix has the form

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{11}(\boldsymbol{\delta}_1) & \Sigma_{12}(\boldsymbol{\delta}_{12}) \\ \Sigma_{12}(\boldsymbol{\delta}_{12}) & \Sigma_{22}(\boldsymbol{\delta}_2) \end{pmatrix}, \tag{4.3}$$

where $\boldsymbol{\delta}_1$, $\boldsymbol{\delta}_2$ and $\boldsymbol{\delta}_{12}$ are different parameter vectors. Then $\dot{\boldsymbol{\Sigma}}_{22}^{\boldsymbol{\delta}_1} = \dot{\boldsymbol{\Sigma}}_{12}^{\boldsymbol{\delta}_1} = \dot{\boldsymbol{\Sigma}}_{11}^{\boldsymbol{\delta}_2} = \dot{\boldsymbol{\Sigma}}_{12}^{\boldsymbol{\delta}_2} = \dot{\boldsymbol{\Sigma}}_{11}^{\boldsymbol{\delta}_{12}} = \dot{\boldsymbol{\Sigma}}_{22}^{\boldsymbol{\delta}_{12}} = \mathbf{0}$ and the score vector is

$$\begin{pmatrix} s_{\boldsymbol{\delta}_1} \\ s_{\boldsymbol{\delta}_2} \\ s_{\boldsymbol{\delta}_{12}} \end{pmatrix} =$$

$$\begin{pmatrix} \frac{1}{2}\sum_{i=1}^{n}\{(\mathbf{y}_{i1}-\boldsymbol{\mu}_{i1})^T\Delta_{i1}^{-1}\mathbf{W}_{i1}\mathbf{Z}_{i1}\dot{\boldsymbol{\Sigma}}_{11}^{\boldsymbol{\delta}_1}\mathbf{Z}_{i1}^T\mathbf{W}_{i1}\Delta_{i1}^{-1}(\mathbf{y}_{i1}-\boldsymbol{\mu}_{i1}) - tr(\mathbf{W}_{0i1}\mathbf{Z}_{i1}\dot{\boldsymbol{\Sigma}}_{11}^{\boldsymbol{\delta}_1}\mathbf{Z}_{i1}^T)\} \\ \frac{1}{2}\sum_{i=1}^{n}\{(\mathbf{y}_{i2}-\boldsymbol{\mu}_{i2})^T\Delta_{i2}^{-1}\mathbf{W}_{i2}\mathbf{Z}_{i2}\dot{\boldsymbol{\Sigma}}_{22}^{\boldsymbol{\delta}_2}\mathbf{Z}_{i2}^T\mathbf{W}_{i2}\Delta_{i2}^{-1}(\mathbf{y}_{i2}-\boldsymbol{\mu}_{i2}) - tr(\mathbf{W}_{0i2}\mathbf{Z}_{i2}\dot{\boldsymbol{\Sigma}}_{22}^{\boldsymbol{\delta}_2}\mathbf{Z}_{i2}^T)\} \\ \sum_{i=1}^{n}\{(\mathbf{y}_{i2}-\boldsymbol{\mu}_{i2})^T\Delta_{i2}^{-1}\mathbf{W}_{i2}\mathbf{Z}_{i2}\dot{\boldsymbol{\Sigma}}_{12}^{\boldsymbol{\delta}_{12}}\mathbf{Z}_{i2}^T\mathbf{W}_{i1}\Delta_{i1}^{-1}(\mathbf{y}_{i1}-\boldsymbol{\mu}_{i1})\} \end{pmatrix}$$

Lin showed that the information matrix in the univariate GLMM depends only on the first two moments of the response variables and its elements can be expressed in closed form for exponential family responses. It is easy to verify that the information matrix for the multivariate GLMM also depends only on the first two moments of the two response variables and does not contain more complicated expressions than its simpler counterpart. The latter property is due to the independence of the response variables under H_0. Note that

$$\mathbf{I}_{\boldsymbol{\delta}\boldsymbol{\delta}} = E(\mathbf{s}_{\boldsymbol{\delta}}\mathbf{s}_{\boldsymbol{\delta}}^T)$$

$$\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\delta}} = E(\mathbf{s}_{\boldsymbol{\beta}}\mathbf{s}_{\boldsymbol{\delta}}^T)$$

$$\mathbf{I}_{\boldsymbol{\beta}\boldsymbol{\beta}} = E(\mathbf{s}_{\boldsymbol{\beta}}\mathbf{s}_{\boldsymbol{\beta}}^T),$$

where the expectations are computed at $\boldsymbol{\delta} = \mathbf{0}$. Consider only $\mathbf{I}_{\boldsymbol{\delta}\boldsymbol{\delta}}$ for now and let us assume that $\boldsymbol{\Sigma}$ has the structure in (4.3). Then $\mathbf{I}_{\boldsymbol{\delta}_1\boldsymbol{\delta}_1}$ is exactly the same as in a univariate GLMM and hence can be expressed as proposed by Lin. On the other hand $\mathbf{I}_{\boldsymbol{\delta}_1\boldsymbol{\delta}_2} = E(\mathbf{s}_{\boldsymbol{\delta}_1}\mathbf{s}_{\boldsymbol{\delta}_2}) = E(\mathbf{s}_{\boldsymbol{\delta}_1})E(\mathbf{s}_{\boldsymbol{\delta}_2}) = 0$ under H_0 because the two score vectors depend on different response variables which are independent under the null hypothesis. Also

$$\mathbf{I}_{\boldsymbol{\delta}_1\boldsymbol{\delta}_{12}} = E(\frac{1}{2}\sum_{i=1}^{n}\{(\mathbf{y}_{i1}-\boldsymbol{\mu}_{i1})^T\Delta_{i1}^{-1}\mathbf{W}_{i1}\mathbf{Z}_{i1}\dot{\boldsymbol{\Sigma}}_{11}^{\boldsymbol{\delta}_1}\mathbf{Z}_{i1}^T\mathbf{W}_{i1}\Delta_{i1}^{-1}(\mathbf{y}_{i1}-\boldsymbol{\mu}_{i1}) - tr(\mathbf{W}_{0i1})\}$$

$$\times\sum_{i=1}^{n}\{(\mathbf{y}_{i1}-\boldsymbol{\mu}_{i1})^T\Delta_{i1}^{-1}\mathbf{W}_{i1}\mathbf{Z}_{i1}\dot{\boldsymbol{\Sigma}}_{12}^{\boldsymbol{\delta}_{12}}\mathbf{Z}_{i2}^T\mathbf{W}_{i2}\Delta_{i2}^{-1}(\mathbf{y}_{i2}-\boldsymbol{\mu}_{i2})\})$$

$$= \sum_{i=1}^{n} \sum_{i'=1}^{n} E[(h_1(\mathbf{y}_{i1})(\mathbf{y}_{i'2} - \boldsymbol{\mu}_{i'2})] =$$

$$\sum_{i=1}^{n} \sum_{i'=1}^{n} Eh_1(\mathbf{y}_{i1}) E(\mathbf{y}_{i'2} - \boldsymbol{\mu}_{i'2}) = \mathbf{0}.$$

Here $h_1(\mathbf{y}_{i1})$ is a function of the first response variable $\mathbf{y}_{i1}$ only. Similarly all other parts of the information matrix which correspond to partial derivatives with respect to parameters for different response variables are zero and hence the expected information matrix has the form

$$\begin{pmatrix} \mathbf{I}_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1} & \mathbf{0} & \mathbf{I}_{\boldsymbol{\beta}_1\boldsymbol{\delta}_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2} & \mathbf{0} & \mathbf{I}_{\boldsymbol{\beta}_2\boldsymbol{\delta}_2} & \mathbf{0} \\ \mathbf{I}_{\boldsymbol{\beta}_1\boldsymbol{\delta}_1} & \mathbf{0} & \mathbf{I}_{\boldsymbol{\delta}_1\boldsymbol{\delta}_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{\boldsymbol{\beta}_2\boldsymbol{\delta}_2} & \mathbf{0} & \mathbf{I}_{\boldsymbol{\delta}_2\boldsymbol{\delta}_2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_{\boldsymbol{\delta}_{12}\boldsymbol{\delta}_{12}} \end{pmatrix}$$

and the score statistic separates as follows

$$\mathbf{s}_{\boldsymbol{\delta}_1}^T (I_{\boldsymbol{\delta}_1\boldsymbol{\delta}_1} - I_{\boldsymbol{\delta}_1\boldsymbol{\beta}_1} I_{\boldsymbol{\beta}_1\boldsymbol{\beta}_1}^{-1} I_{\boldsymbol{\beta}_1\boldsymbol{\delta}_1})^{-1} \mathbf{s}_{\boldsymbol{\delta}_1}$$

$$+ \mathbf{s}_{\boldsymbol{\delta}_2}^T (I_{\boldsymbol{\delta}_2\boldsymbol{\delta}_2} - I_{\boldsymbol{\delta}_2\boldsymbol{\beta}_2} I_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2}^{-1} I_{\boldsymbol{\beta}_2\boldsymbol{\delta}_2})^{-1} \mathbf{s}_{\boldsymbol{\delta}_2}$$

$$+ \mathbf{s}_{\boldsymbol{\delta}_{12}}^T I_{\boldsymbol{\delta}_{12}\boldsymbol{\delta}_{12}}^{-1} \mathbf{s}_{\boldsymbol{\delta}_{12}}.$$

This factorization appears only when the variance-covariance matrix is structured as in (4.3), otherwise the expression is more complicated but still depends only on the first two moments of the response. The key to proving this is to notice that the highest order expectation that needs to be computed is of the form $E(y_{ij} - \mu_{ij})^4$. The same is true for the univariate GLMM and hence Lin's arguments can be directly applied.

Lin proves that the global score statistic in the univariate GLMM follows a chi-squared distribution with d degrees of freedom (d is equal to the number of random effects) asymptotically under $\boldsymbol{\delta} = \mathbf{0}$. The asymptotic result holds when the number

of subjects goes to infinity and the number of observations on each subjects remains bounded. In the multivariate GLMM the asymptotic distribution is also χ^2 but with the number of degrees of freedom adjusted accordingly.

The global score statistic is not very useful even in the GLMM contest because it tests the significance of all variance components simultaneously, while in most cases it will be more interesting to check a subset of the variance components. But in the multivariate GLMM model the global score test is even less appealing. Suppose that the test is performed and that the null hypothesis is rejected. What information can one get from that result? It will not be clear whether the rejection occured because of extra variability in one of the variables, or in the other one, or in both. It may be more meaningful to perform score tests for each variable separately and then check for correlation between the two responses.

Lin also develops score tests for specific variance components in the independent random effects model. In contrast to the global test, here the score vector and the efficient information matrix can not be computed in closed form in general and Lin uses Laplace approximations. Not surprisingly, the approximation to the score statistic does not work well in the binary case as demonstrated by some simulations that she performed. Lin's score tests can be used with the Breslow and Clayton, and the Wolfinger and O'Connell methods but are not adequate if used with Gaussian quadrature or the MCEM algorithm. In that case it is natural to try to develop score tests based on numerical or stochastic approximations. We now discuss a direct approach which is of limited use, and an indirect approach which is more complicated but is especially suited for variance components on the boundary of the parameter space.

Tests for Individual Variance Components

We consider a bivariate GLMM for simplicity. Suppose one is interested in testing $H_0 : \boldsymbol{\psi}_l = \mathbf{0}$, where $\boldsymbol{\psi}_l$ is a subset of the variance components for the random effects.

Also let ψ_l have L elements and let $\psi = (\psi_l^T, \psi_{-l}^T)^T$. The score statistic is

$$T_s = \mathbf{s}_{\psi_l}^T (\mathbf{I}_{\psi_l\psi_l} - \mathbf{I}_{\psi_l\psi_{-l}} \mathbf{I}_{\psi_{-l}\psi_{-l}}^{-1} \mathbf{I}_{\psi_{-l}\psi_l})^{-1} \mathbf{s}_{\psi_l}.$$

$$T_s \overset{H_0}{\sim} \chi_L^2$$

To develop a Monte Carlo approximation we can try to follow the approach we used for the conditional independence test. Under the assumption of interchangeability of the integral and differential signs the score vector can be rewritten as follows

$$\mathbf{s}_{\psi_l} = \sum_{i=1}^{n} E(\frac{\partial}{\partial \psi_l} ln f(\mathbf{b}_i, \mathbf{\Sigma}) | \mathbf{y}_i, \psi).$$

The random effects density $f(\mathbf{b}_i, \boldsymbol{\delta})$ is multivariate normal for our models and hence it is possible to obtain expressions for the partial derivatives inside the expectation using the approach of Jennrich and Schluster as outlined in Section 3.2.1. There is no closed form expression for the score vector but we can approximate it by

$$\frac{1}{m} \sum_{i=1}^{n} \sum_{k=1}^{m} \frac{\partial}{\partial \psi_l} ln f(\mathbf{b}_i^{(k)}, \tilde{\psi}_{-l}),$$

where $\mathbf{b}_i^{(k)}$ are simulated values from the conditional distribution $\mathbf{b}_i | \mathbf{y}_i$. As mentioned before the expected information matrix is much harder to work with and hence the observed information matrix can be approximated using Louis' method as shown in Section 3.2.2. Notice though that the score vector must be evaluated at $\psi_l = \mathbf{0}$ and at the restricted maximum likelihood estimates $\tilde{\psi}_{-l}$. Depending on the subset of the variance components tested, the derivative at $\psi_l = \mathbf{0}$ may not exist. Consider for example the case when $\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$ and when the null hypothesis is H_0 : $\sigma_1^2 = \sigma_{12} = 0$. Then $\mathbf{\Sigma}$ is singular under H_0 and we can not evaluate the derivative. If we test only $H_0 : \sigma_{12} = 0$ then there is no problem with the test. In this case

the parameter is not on the boundary under the null hypothesis, but the score test is useful because the correlation between the response variables can be tested from the fit of two separate GLMMs. Recall that univariate GLMMs can be fitted using standard software such as PROC NLMIXED in SAS. Hence one can decide whether there is a need to fit the two responses together based only on univariate analyses.

An alternative method to compute an approximation to the score statistic above is to use Gauss-Hermite quadrature. Two possible approaches can be followed. The easier one is to compute first and second order numerical derivatives of the log-likelihood and then compute the score statistic based on them. Exact derivatives might be useful if the number of observations per subject is not very large. Both approaches are not applicable if the tested subset of variance components leads to a nonpositive-definite $\mathbf{\Sigma}$.

Denote the Gauss-Hermite quadrature approximation of the log-likelihood by l^{GQ}. Also let $\mathbf{s}^{GQ}_{\boldsymbol{\psi}_l} = \frac{\partial l^{GQ}}{\partial \boldsymbol{\psi}_l}$ and $\mathbf{J}^{GQ}_{\boldsymbol{\psi}_l,\boldsymbol{\psi}_{-l}} = -\frac{\partial^2 l^{GQ}}{\partial \boldsymbol{\psi}_l \partial \boldsymbol{\psi}_{-l}}$, i.e the score and the observed information matrix are approximated by the first and second order numerical derivatives of the Gauss-Hermite approximation to the log-likelihood. The score statistic for testing $H_0 : \boldsymbol{\psi}_l = \mathbf{0}$ then is approximated by

$$\mathbf{s}^{GQ^T}_{\boldsymbol{\psi}_l}(\mathbf{J}^{GQ}_{\boldsymbol{\psi}_l,\boldsymbol{\psi}_l} - \mathbf{J}^{GQ}_{\boldsymbol{\psi}_l,\boldsymbol{\psi}_{-l}}(\mathbf{J}^{GQ}_{\boldsymbol{\psi}_{-l},\boldsymbol{\psi}_{-l}})^{-1}\mathbf{J}^{GQ}_{\boldsymbol{\psi}_{-l},\boldsymbol{\psi}_l})^{-1}\mathbf{s}^{GQ}_{\boldsymbol{\psi}_l},$$

where the score and the information matrix are evaluated at $\boldsymbol{\psi}_l = 0$ and at the restricted maximum likelihood estimates $\tilde{\boldsymbol{\psi}}_{-l}$.

As mentioned before this direct approach to testing for variance components will not work in many interesting cases when the parameters are on the boundary of the paramater space. To be able to handle such a problem the integrand should be modified to avoid dependence on the random effects with 0 variances. Notice that if we want to test whether for example $\sigma^2_k = 0$ we have to set all corresponding covariance terms $\sigma_{kk'}$ to be equal to 0 as well. Suppose we want to test $\boldsymbol{\psi}_l = 0$ and

let the corresponding vectors of random effects with 0 variances be $\mathbf{b}_i^l$. Let

$$\mathbf{b}_i = \begin{pmatrix} \mathbf{b}_i^l \\ \mathbf{b}_i^{-l} \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}^{l,l} & \boldsymbol{\Sigma}^{l,-l} \\ \boldsymbol{\Sigma}^{-l,l} & \boldsymbol{\Sigma}^{-l,-l} \end{pmatrix}.$$

Under H_0 $\boldsymbol{\Sigma}^{l,l} = \mathbf{0}$ and $\boldsymbol{\Sigma}^{l,-l} = \mathbf{0}$.

We can rewrite the marginal likelihood for the i^{th} subject as follows

$$f_i(\mathbf{y}_i; \boldsymbol{\psi}) = E_{\mathbf{b}_i^{-l}} E_{\mathbf{b}_i^l|\mathbf{b}_i^{-l}}[exp(l_i(\mathbf{b}_i))]$$

and then we can expand the integrand around the conditional mean of $\mathbf{b}_i^l$. Because the random effects distribution is assumed to be multivariate normal, the conditional distribution of $\mathbf{b}_i^l$ is also multivariate normal

$$\mathbf{b}_i^l|\mathbf{b}_i^{-l} \sim N(\boldsymbol{\Sigma}^{l,-l}(\boldsymbol{\Sigma}^{-l,-l})^{-1}\mathbf{b}_i^{-l}, \boldsymbol{\Sigma}_{ll}), \quad \boldsymbol{\Sigma}_{ll} = \boldsymbol{\Sigma}^{l,l} - \boldsymbol{\Sigma}^{l,-l}(\boldsymbol{\Sigma}^{-l,-l})^{-1}\boldsymbol{\Sigma}^{-l,l}.$$

Thus we obtain $E_{\mathbf{b}_i^l|\mathbf{b}_i^{-l}} exp(l_i(\mathbf{b}_i)) =$

$$exp(l_i(\mathbf{b}_i^{-l}))[1 + \frac{1}{2}tr(\mathbf{Z}_i^{lT}(\frac{\partial l_i(\mathbf{b}_i^{-l})}{\partial \boldsymbol{\eta}_i}\frac{\partial l_i(\mathbf{b}_i^{-l})}{\partial \boldsymbol{\eta}_i^T} + \frac{\partial^2 l_i(\mathbf{b}_i^{-l})}{\partial \boldsymbol{\eta}_i \partial \boldsymbol{\eta}_i^T})\mathbf{Z}_i^l\boldsymbol{\Sigma}_{ll}) + r_i], \tag{4.4}$$

where $\mathbf{Z}_i^l$ contains the rows of $\mathbf{Z}_i$ corresponding to $\mathbf{b}_i^l$, and $l_i(\mathbf{b}_i^{-l})$ denotes the conditional log-likelihood for the i^{th} subject $lnf(\mathbf{y}_i|\mathbf{b}_i, \boldsymbol{\psi})$ evaluated at $\mathbf{b}_i^{-l}$ and at $\mathbf{b}_i^l = \boldsymbol{\Sigma}^{l,-l}(\boldsymbol{\Sigma}^{-l,-l})^{-1}\mathbf{b}_i^{-l}$. To obtain the marginal likelihood of $\mathbf{y}_i$ we take the expectation of (4.4) with respect to $\mathbf{b}_i^{-l}$ and because there is no closed form expression for the integral, we need to use approximations. We ignore the remainder term r_i which depends on second and higher order products of variance components which will be 0 under H_0. Now the problem of approximating the score statistic using the indirect approach is the same as the one using the direct approach but with $exp(l_i(\mathbf{b}_i))$ in the integrand replaced by (4.4) and the random vector $\mathbf{b}_i$ replaced by $\mathbf{b}_i^{-l}$. Hence Gaussian quadrature and Monte Carlo methods can be used as described earlier in this section.

4.4 Applications

In this chapter we consider numerical and stochastic approximations to the Wald, likelihood ratio and score tests. The conditional independence test provides a nice framework to compare the performance of all three approaches. Recall that in the ethylene glycol example conditional independence implies that the correlation between birth weight and malformation measured on the same fetus is the same as the correlation between birth weight and malformation measured on two different fetuses within a litter. Hence it is likely that this assumption will not be satisfied. To check the assumption we specify a more complicated model as suggested in Section 4.3.2.

y_{i1j} - fetal weight of j^{th} fetus in i^{th} litter

y_{i2j} - malformation status of j^{th} fetus in i^{th} litter

d_i - dose administered to i^{th} litter

$y_{i1j}|\mathbf{y}_{i2}, b_{i1} \sim$ *indep* $N(\mu_{i1j}, \sigma^2)$

$y_{i2j}|b_{i2} \sim$ *indep* $Be(\mu_{i2j})$

$\mu_i 1j = \beta_{10} + \beta_{11} d_i + \gamma y^*_{i2j} + b_{i1}$, where $y^*_{i2j} = 2y_{i2j} - 1$

$logit(\mu_{i2j}) = \beta_{20} + \beta_{21} d_i + b_{i2}$

$\mathbf{b}_i = (b_{i1}, b_{i2})^T \sim N_2(0, \mathbf{\Sigma})$, $\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$.

The score test statistic for $\gamma = 0$ using 100 quadrature points is 17.45 ($p-value < 0.0001$) and therefore the hypothesis of conditional independence is rejected. Hence the model introduced above may be more appropriate for the ethylene glycol data. We fit this model in order to compare the score, Wald and likelihood-ratio statistics. The estimates using Gaussian quadrature and the Monte Carlo EM algorithm for logit and for probit links are given in Table 4.1. Gaussian quadrature used 100 quadrature points both for the logit and for the probit model. It took 3 hrs 40 min for the logit models and 5 hrs and 30 min for the probit model to converge. The numbers of

iterations were 32 and 50 respectively. The MCEM algorithms took about 15 hrs to converge: the logit model used 63 iterations and had a final sample size of 7898, the probit model used 52 iterations with a final sample size of 5549. The starting values were the same for all algorithms: gamma had an initial value of zero and all the other parameters had initial values as in the models without gamma (Section 3.4).

The estimates of γ in all fits were identical up to three places after the decimal point and the standard errors were also very similar. There is evidence that the fetal weight of a fetus significantly decreases if malformation status is changed from absent to present. Notice that in this example there is no natural ordering of the response variables and hence the decision to include malformation in the linear predictor for fetal weight is for mathematical convenience. Because of the identity link function for the normal response the expression for the Monte Carlo approximation of the score vector is more simple. Also, the interpretation of the γ coefficient is easier to understand. In that case 2γ is interpreted as the amount by which a subject's fetal weight is expected to decrease if malformation status is changed from absence to presence. In contrast, if we were to include fetal weight in the linear predictor for malformation and we used a logit link, then the interpretation of γ would be the change in a subject's log odds for malformation per unit change in birth weight, controlling for dose.

To compare the performance of the stochastic and analytical approximations to the test statistic we computed Gaussian quadrature and Monte Carlo approximations using different numbers of quadrature points and simulation sample sizes. Attention was restricted to the logit link and we used the final parameter estimates from the Gaussian quadrature fits for the conditional independence (Section 3.4) and conditional dependence models.

The results are provided in Tables 4.2 and 4.3. As the number of quadrature points increases the Gauss-Hermite approximations improve, or at least they are internally

Table 4.1. Estimates from the conditional dependence model fit to the ethylene glycol data

Par.	GQ(logit)		MCEM(logit)		GQ (probit)		MCEM (probit)	
	Est	SE	Est	SE	Est	SE	Est	SE
β_{10}	0.936	0.014	0.937	0.024	0.936	0.014	0.937	0.027
β_{11}	-0.081	0.008	-0.082	0.010	-0.081	0.008	-0.082	0.010
γ	-0.016	0.004	-0.016	0.004	-0.016	0.004	-0.016	0.005
β_{20}	-4.371	0.421	-4.403	0.537	-2.420	0.218	-2.444	0.300
β_{21}	1.768	0.212	1.782	0.238	0.981	0.113	0.991	0.120
σ	0.074	0.002	0.074	0.002	0.074	0.002	0.074	0.002
σ_1	0.083	0.007	0.083	0.007	0.083	0.007	0.083	0.007
σ_2	1.534	0.201	1.522	0.206	0.851	0.109	0.844	0.112
ρ	-0.612	0.100	-0.615	0.102	-0.594	0.101	-0.599	0.103

Table 4.2. Gaussian quadrature approximations to the score, Wald and likelihood ratio statistics for testing for conditional independence in the ethylene glycol example

Number of quadrature points	score	Wald	LR
10	12.10	11.07	19.65
20	19.33	18.54	17.37
30	-63.44	17.60	18.09
40	16.58	17.09	16.97
50	17.20	17.05	16.95
60	17.32	17.04	16.92
70	17.40	17.06	16.95
80	17.43	17.07	16.96
90	17.45	17.08	16.96
100	17.45	17.13	16.96

consistent and show decreasing variability (Table 4.2). The only negative estimate for the score statistic is due to a non-positive definite estimate of the observed information matrix. As pointed out before there is no guarantee that this will not happen. It seems that about 50 quadrature points are adequate to approximate the test statistics in this example.

It is not surprising that the Monte Carlo approximations show more variability (Table 4.3). We use two different random seeds and hence for each test there are two columns of values. Of the three test statistics the likelihood ratio shows most variability. It is based on an approximation of the log-likelihood rather than on an

Table 4.3. Monte Carlo approximations to the score (S), Wald (W) and likelihood ratio (LR) statistics for conditional independence in the ethylene glycol example. Two different initial random seeds are used.

Sample size	S1	S2	W1	W2	LR1	LR2
100	13.56	13.85	16.44	15.92	17.84	5.17
500	18.73	29.26	15.90	16.02	17.36	13.80
1000	16.18	17.95	17.24	21.37	12.44	19.15
5000	17.24	17.02	16.69	16.67	18.75	18.28
10000	17.14	16.95	16.77	17.17	18.85	17.32
20000	17.09	17.07	17.23	16.90	16.08	17.10

approximation of the information matrix which may indicate that larger sample sizes are needed to approximate the log-likelihood precisely than are needed for the information matrix. A reasonable simulation sample size to use in view of computational efficiency is 5000. Note that rejection sampling is not used for the likelihood ratio statistic so the approximation takes less time.

In the next section we further study the performance of the score, Wald and likelihood ratio statistics for conditional independence via a simulation study.

4.5 Simulation Study

We use the structure of the simulated data example (Section 3.3) with either 30 or 100 subjects and 10 bivariate observations per subject. The parameters are as follows: $\beta_1 = 4$, $\beta_2 = 1$, $\sigma^2 = 1$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\sigma_{12} = 0.5$ and five different values of γ: $0, 0.05, 1, 2$ and 3. $\gamma = 0$ corresponds to conditional independence. A bivariate GLMM and a bivariate GLMM with conditional dependence are fitted using Gaussian quadrature with 50 quadrature points. Approximations to the Wald, likelihood ratio and score statistics for testing $\gamma = 0$ are computed for 50 quadrature points and for Monte Carlo simulation sample sizes of 5000. 50 samples are generated, and means, standard deviations and rejection percentages are computed for all three statistics.

The results using Gaussian quadrature are summarized in Tables 4.4 - 4.7, and those using Monte Carlo approximations are summarized in Tables 4.8 - 4.11. Under

$H_0 : \gamma = 0$, the mean and the standard deviations of the test statistics should be 1 and $\sqrt{2} = 1.41$ respectively. For $\gamma = 0$ for the larger simulation sample all three Gaussian quadrature approximations give a mean of 0.91 and a standard deviation of about 1.37. The corresponding Monte Carlo approximations to the Wald and score tests gives essentially the same means and slightly lower (1.34 as compared to 1.37) standard deviations. Only the Monte Carlo approximations to the likelihood ratio statistic has bigger mean value and bigger standard deviation. For the smaller sample size almost all values (except the mean of the Monte Carlo likelihood ratio statistic) are further away from the truth: about 0.86 and 1.10. We expect those values to be closer to the truth if the simulation sample size is increased. Interestingly, for all settings using Gaussian quadrature the score statistic has the largest average value and the likelihood ratio statistic has the smallest. No such trend is obvious in the Monte Carlo approximations. The differences in the results for the Monte Carlo likelihood ratio statistic may be attributed to the different kind of approximation used as compared to the approximation of the information matrix needed for the Wald and score statistic.

Another way to summarize the results is to look at the percentage of simulations in which the null hypothesis is rejected for different γ levels. These are given in Tables 4.5 and 4.7 for the Gaussian quadrature approximations, and in Tables 4.9 and 4.11 for the Monte Carlo approximations. From that perspective the three statistics are almost identical and their type I error rates are close to the nominal levels. (Recall that the simulation sample size is only 50 and that is why only certain percentages could be observed. The simulation sample size was chosen for reasons of computational feasibility). Among the Gaussian quadrature approximations usually the score statistic has the highest rejection rate and the only case when it rejects less often is in the smaller sample setting when $\gamma = 0.3$. This is due to one sample in which

Table 4.4. Means and standard deviations of the Gaussian quadrature approximations to the score, Wald and likelihood ratio test statistics for conditional independence: sample size = 100 subjects

	score		Wald		likelihood ratio	
γ	mean	s.d.	mean	s.d.	mean	s.d.
0	0.91	1.37	0.91	1.37	0.91	1.36
0.05	2.00	2.43	1.99	2.41	1.98	2.40
0.1	8.48	4.68	8.38	4.57	8.33	4.52
0.2	27.99	10.57	27.03	9.94	26.59	9.66
0.3	64.13	16.85	59.62	14.56	57.64	13.61

the estimate of the variance-covariance matrix in the reduced model was non-positive definite and the score statistic was set to be equal to zero.

Among the Monte Carlo approximations, the Wald and score statistics show almost perfect agreement with each other. The only exception is the small sample setting when $\gamma = 0.3$ for the same reasons as discussed above. The Monte Carlo likelihood ratio statistic shows more variability and may require larger simulation sample size. The Gaussian quadrature and Monte Carlo approximations of the Wald and score test statistics are similar with the Monte Carlo approximations having slightly lower standard deviations.

This limited simulation study indicates that the Gaussian quadrature approximations of all three statistics, and the Monte Carlo approximations of the Wald and score statistics perform similarly and the choice of which one to use should probably be dictated by other practical considerations. For example, if fitting a more complicated model is computationally intensive probably the score statistic should be used. If the extended model is going to be fit anyway, the least computationally demanding statistic is the likelihood ratio statistic. But if a Monte Carlo approximation is preferred for the likelihood ratio statistic we may need a larger simulation sample size which may obscure the computational advantage. Further study is needed to determine whether the observed behaviour is typical of the approximated test statistics or dictated by the particular setting under consideration.

Table 4.5. Rejection rates for the Gaussian quadrature approximations to the score, Wald and likelihood ratio test statistics for conditional independence: sample size = 100 subjects

	$\alpha = 0.01$			$\alpha = 0.05$			$\alpha = 0.10$		
	S	W	LR	S	W	LR	S	W	LR
γ	%	%	%	%	%	%	%	%	%
0	0	0	0	0.06	0.06	0.06	0.08	0.08	0.08
0.05	0.06	0.06	0.04	0.20	0.18	0.18	0.30	0.30	0.30
0.1	0.64	0.62	0.62	0.86	0.86	0.84	0.92	0.92	0.92
0.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

S: score, W: Wald, LR: likelihood ratio

Table 4.6. Means and standard deviations of the Gaussian quadrature approximations to the score, Wald and likelihood ratio test statistics for conditional independence: sample size = 30 subjects

	score		Wald		likelihood ratio	
γ	mean	s.d.	mean	s.d.	mean	s.d.
0	0.87	1.11	0.86	1.10	0.86	1.09
0.05	1.42	1.92	1.40	1.87	1.39	1.84
0.1	3.24	3.41	3.16	3.26	3.12	3.19
0.2	8.42	5.28	8.07	4.91	7.91	4.74
0.3	19.65	8.96	18.24	7.38	17.56	6.86

Table 4.7. Rejection rates for Gaussian quadrature approximations to the score, Wald and likelihood ratio test statistics for conditional independence: sample size = 30 subjects

	$\alpha = 0.01$			$\alpha = 0.05$			$\alpha = 0.10$		
	S	W	LR	S	W	LR	S	W	LR
γ	%	%	%	%	%	%	%	%	%
0	0	0	0	0.02	0.02	0.02	0.10	0.10	0.10
0.05	0.04	0.04	0.04	0.10	0.10	0.08	0.16	0.14	0.14
0.1	0.12	0.12	0.12	0.30	0.30	0.30	0.46	0.44	0.44
0.2	0.58	0.56	0.56	0.82	0.82	0.82	0.86	0.86	0.86
0.3	0.96	0.98	0.98	0.98	1.00	1.00	0.98	1.00	1.00

S: score, W: Wald, LR: likelihood ratio

Table 4.8. Means and standard deviations of the Monte Carlo approximations to the score, Wald and likelihood ratio test statistics for conditional independence: sample size = 100 subjects

	score		Wald		likelihood ratio	
γ	mean	s.d.	mean	s.d.	mean	s.d.
0	0.90	1.34	0.90	1.34	1.07	2.00
0.05	1.98	2.37	1.98	2.37	2.00	2.78
0.1	8.35	4.48	8.35	4.52	8.24	4.43
0.2	26.55	9.47	26.92	9.78	26.21	10.09
0.3	57.80	13.56	59.61	14.51	57.56	13.79

Table 4.9. Rejection rates for the Monte Carlo approximations to the score, Wald and likelihood ratio test statistics for conditional independence: sample size = 100 subjects

	$\alpha = 0.01$			$\alpha = 0.05$			$\alpha = 0.10$		
	S	W	LR	S	W	LR	S	W	LR
γ	%	%	%	%	%	%	%	%	%
0	0	0	0	0.06	0.06	0.10	0.08	0.08	0.12
0.05	0.04	0.04	0.06	0.20	0.20	0.20	0.30	0.30	0.26
0.1	0.64	0.62	0.58	0.86	0.86	0.84	0.92	0.92	0.96
0.2	1.00	1.00	0.96	1.00	1.00	0.98	1.00	1.00	0.98
0.3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

S: score, W: Wald, LR: likelihood ratio

Table 4.10. Means and standard deviations of the Monte Carlo approximations to the score, Wald and likelihood ratio test statistics for conditional independence: sample size = 30 subjects

	score		Wald		likelihood ratio	
γ	mean	s.d.	mean	s.d.	mean	s.d.
0	0.86	1.09	0.86	1.09	0.96	1.10
0.05	1.39	1.84	1.40	1.87	1.40	2.01
0.1	3.12	3.14	3.16	3.25	3.00	3.25
0.2	7.91	4.72	8.11	4.94	7.93	4.69
0.3	17.81	6.91	18.18	7.42	17.60	7.07

Table 4.11. Rejection rates for the Monte Carlo approximations to the score, Wald and likelihood ratio test statistics for conditional independence: sample size = 30 subjects

	$\alpha = 0.01$			$\alpha = 0.05$			$\alpha = 0.10$		
	S	W	LR	S	W	LR	S	W	LR
γ	%	%	%	%	%	%	%	%	%
0	0	0	0	0.02	0.02	0.04	0.10	0.10	0.08
0.05	0.04	0.04	0.04	0.10	0.08	0.08	0.16	0.16	0.20
0.1	0.12	0.12	0.14	0.30	0.30	0.28	0.44	0.44	0.44
0.2	0.56	0.56	0.50	0.82	0.82	0.86	0.86	0.86	0.90
0.3	0.96	0.98	0.98	0.98	1.00	1.00	0.98	1.00	1.00

S: score, W: Wald, LR: likelihood ratio

4.6 Future Research

An important future research topic is to compare the proposed numerical and stochastic approximations of the variance component score statistics (Section 4.3.3) to the analytical approximations developed by Lin. The expectations are that the Gauss-Hermite quadrature and Monte Carlo methods will be more computationally intensive but will outperform the Laplace approximation methods when the data is far from normally ditributed.

It is also of interest to consider tests for conditional independence for different settings in addition to the Bernoulli-Normal combination discussed here. Alternative approaches to incorporate dependence in the model should also be investigated. One such approach is introduced in the next chapter but it is applicable only to continuous and binary variables. What can be done in an application like the pony data is not clear.

In this chapter we did not address some important research questions such as model goodness-of-fit and effects of departures from the parametric assumptions on the estimates. In the case of dicrete response variables it may be possible to use the deviance statistic to check the model fit. But when the data is continuous, or

both discrete and continuous, this question becomes much more complicated. The effects of departures from the parametric assumptions on the maximum likelihood estimates can probably be studied in the framework proposed by White (1982). He showed that the maximum likelihood estimates under a false model converge to a value which minimizes the Kullback-Leibler divergence between the true and misspecified models. Some approximations will need to be used to assess the magnitude of the introduced bias.

Finally, the challenges of verifying the consistency and asymptotic normality of the maximum likelihood estimates in the GLMM and of quantifying the precision of their numerical and stochastic approximations are still unresolved and require further investigation.

CHAPTER 5
CORRELATED PROBIT MODEL

Using the GLMM for multivariate repeated measures allows for modelling of any mixture of outcomes in the exponential family but requires the rather restrictive assumption of conditional independence between the responses given the random effects. It is difficult to construct a general fully parametric model that overcomes this drawback because of the need to define multivariate distributions for mixtures of responses. However, in the special case of a mixture of one binary and one continuous response (as in the ethylene glycol example), one can fit a correlated probit model with an underlying latent variable for the binary response. Catalano and Ryan (1992) considered such a model and used GEE methodology for estimation. In this section we propose a Monte Carlo EM algorithm for finding 'exact' ML estimates. Chan and Kuk (1997) introduced such an algorithm for models with binary responses, but as we will show it can also be used for a mixture of binary and normal variables and in the case of correlated errors. We also consider an acceleration to this algorithm using modifications proposed by Liao (1999) and by Lavielle, Delyon and Moulines (1999).

The introduction to this chapter contains a literature overview of latent variable models for binary, ordinal and censored continuous data. First, we consider models for cross-sectional data and then we mention some extensions to correlated data. Section 5.2 defines the correlated probit model and Section 5.3 contains a description of the model fitting method. Results from the analysis of the ethylene glycol example are presented in Section 5.4. Section 5.5 is devoted to a simulation study investigating efficiency gains in the correlated probit model, and an identifiability issue is discussed

in Section 5.6. Section 5.7 describes the extensions of this model to any mixture of binary, continuous, continuous censored and ordinal data with known thresholds and the chapter concludes with a discussion of future research directions (Section 5.8).

5.1 Introduction

In many applications an observed binary variable y can be assumed to result from a dichotomization of an unobserved (latent) continuous variable y^* ranging from $-\infty$ to $+\infty$. Larger values of y^* are observed as $y = 1$ while smaller values of y^* are observed as $y = 0$. Motivation for this representation often comes from studies of dose-response relationships in populations of biological organisms where the response of interest is whether a randomly selected individual receiving a certain dose of a toxic chemical dies (Finney, 1964; Ashford and Sowden, 1970). The latent variable y^* is assumed to be linearly related to the observed covariates through the model

$$y_i^* = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \tag{5.1}$$

where ϵ_i are i.i.d. with some continuous distribution. The observed binary variable is linked to the latent continuous variable in the following way:

$$y_i = \begin{cases} 1 & if \quad y_i^* > \tau \\ 0 & if \quad y_i^* \leq \tau \end{cases}$$

In many applications $\epsilon_i \sim N(0, \sigma^2)$ and the probability $p = P(y_i = 1)$ is

$$P(y_i^* > \tau) = P(\mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i > \tau) = P(\epsilon_i > \tau - \mathbf{x}_i^T \boldsymbol{\beta}) = 1 - \Phi(\frac{\tau - \mathbf{x}_i^T \boldsymbol{\beta}}{\sigma}) = \Phi(\frac{\mathbf{x}_i^T \boldsymbol{\beta} - \tau}{\sigma}).$$

Without loss of generality the threshold τ can be set equal to zero because it can be absorbed in the intercept of the linear predictor ($\beta_0^* = \beta_0 - \gamma$). The error variance σ^2 is not estimable from the data because y_i^* is not observed and it is only known whether y_i^* is positive or negative. Usually σ^2 is taken to equal 1.

The normal distribution assumption for the underlying latent variable leads to a probit model for the binary response:

$$\Phi^{-1}(p) = \mathbf{x}_i^T\boldsymbol{\beta},$$

where Φ is the cumulative density funstion of a standard normal random variable. Gaddum (1933) and Bliss (1934a, 1934b, 1935a, 1935b) are credited with the development of the method of probit analysis.

Ordinal data are also often assumed to arise from an underlying latent continuous variable. Let y_i^* be related to covariates as described in (5.1) and let $y_i = r$ if $\tau_{r-1} \leq y_i^* < \tau_r$ for $r = 1, 2, ...R$. If the errors are assumed to be distributed $N(0, \sigma^2)$

$$P(y_i = r) = P(\tau_{r-1} \leq y_i^* < \tau_r) = \Phi(\frac{\tau_{r-1} - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}) - \Phi(\frac{\tau_r - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}),$$

$$P(y_i = 1) = \Phi(\frac{\tau_1 - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}),$$

$$P(y_i = R) = 1 - \Phi(\frac{\tau_R - \mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}).$$

This ordered probit model was first suggested by Aitchison and Silvey (1957) who considered only a single independent variable. McKelvey and Zavoina (1975) extended Aitchison and Silvey's work to the case of multiple independent variables.

For identifiability reasons usually the thresholds are reparametrized as follows $\tau_1 = 0$, $\tau_2^* = \tau_2 - \tau_1$, ..., $\tau_R^* = \tau_R - \tau_{R-1}$ and σ^2 is set equal to 1.0. In some applications the actual thresholds may be known. For example the response may be family income, but it may only be known what tax bracket the income falls in and not exactly how much it is. This latter case will be of interest in Section 5.7 where extensions of the correlated probit model are considered.

An underlying latent variable is also appropriate for censored continuous data. When the censoring is on the left, observations at or below a certain value are set to

some predefined number. We can again assume an underlying regression model as in (5.1). The observed censored variable is defined as follows:

$$y_i = \begin{cases} y_i^* & if \quad y_i^* > \tau \\ \tau_y & if \quad y_i^* \leq \tau \end{cases}$$

If we again assume that the errors have $N(0, \sigma^2)$ distribution then the resulting model for the observed response is called the "tobit" model. The name stands for 'Tobin's probit' model in honor of Tobin's (1958) work on household expenditures for durable goods which led to the introduction of the tobit model.

Censoring from above, and from above and below simultaneously can also occur. In general $\tau \neq \tau_y \neq 0$ but they are known.

The tobit and probit models are related in the following way: in the tobit model we know the value of y^* when $y^* > \tau$ while in the probit model we only know if $y^* > \tau$. The derivation of the probability of a case being censored is very similar to the derivation of the probability of an event in the probit model. In general the estimates for β from the tobit model are more efficient than the estimates that would be obtained from a probit model, and σ^2 can be estimated from the tobit but not from the probit model.

All three models introduced so far can be fitted using maximum likelihood. Long (1997) provides details about the appropriate iterative algorithms.

Extensions of the above methods for multiple response variables or clustered data have been considered in the literature. Ashford and Sowden (1970) introduced a multivariate probit model based on an underlying multivariate normal distribution. In the bivariate case there are two correlated underlying latent variables $(y_{i1}^*, y_{i2}^*)^T \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \begin{Bmatrix} 1 & \rho \\ \rho & 1 \end{Bmatrix}$, $y_{i1} = I\{y_{i1}^* > 0\}$ and $y_{i2} = I\{y_{i2}^* > 0\}$. Then

$$P(y_{i1} = 0, y_{i2} = 0) = P(y_{i1}^* < 0, y_{i2}^* < 0) = \Phi^{(2)}(-\mu_1, -\mu_2),$$

where $\Phi^{(2)}$ denotes the bivariate standard normal c.d.f. Also

$$P(y_{i1}=0,y_{i2}=1)=P(y_{i1}^{*}<0,y_{i2}^{*}>0)=P(y_{i1}^{*}<0)-P(y_{i1}^{*}<0,y_{i2}^{*}<0)=$$

$$\Phi(-\mu_1)-\Phi^{(2)}(-\mu_1,-\mu_2),$$

and the remaining joint probabilities can be similarly determined.

Ochi and Prentice (1984) introduced a correlated generalization of the multivariate probit model of Ashford and Sowden to fit regression models to exchangeable binary data. The equicorrelated variance-covariance structure $\mathbf{\Sigma}=\sigma^2\{(1-\rho)\mathbf{I}+\rho\mathbf{J}\}$ of the underlying normal distribution allows using approximations to equicorrelated normal integrals to simplify maximum likelihood estimation. Regan and Catalano (1999) considered a generalization of Ochi and Prentice's method for clustered binary and continuous outcomes.

Chan and Kuk (1997) proposed a probit-normal model for binary data with correlated random effects, which provides a great deal of flexibility in modelling diverse correlation structures. The general form of their model is

$$\Phi^{-1}(\mathbf{p})=(\Phi^{-1}(p_1),...,\Phi^{-1}(p_N))^T=\mathbf{X}\boldsymbol{\beta}+\sum_{r=1}^{R}\mathbf{Z}_r\mathbf{b}_r,$$

where $p_j=Pr(y_j=1)$, y_1, y_2, ...y_N are the observed binary variables, $\mathbf{X}$ is an $N\times p$ model matrix, $\boldsymbol{\beta}$ is a $p\times 1$ vector of fixed effects, $\mathbf{b}_r$ is a $q_rk_r\times 1$ vector of random effects with corresponding $N\times q_rk_r$ model matrix $\mathbf{Z}_r$. It is assumed that $\mathbf{b}_1,...\mathbf{b}_k$ are independent and

$$\mathbf{b}_r=\begin{pmatrix}\mathbf{b}_{r1}\\ \vdots\\ \mathbf{b}_{rq_r}\end{pmatrix}\sim N_{q_r\times k_r}(0,I_{q_r}\otimes\Sigma_r).$$

Chan and Kuk viewed the above probit-linear mixed model as a threshold model resulting from dichotomizing the observations from a Gaussian mixed model. In other

words, they assumed that $y_j = I\{y_j^* > 0\}$ and

$$\mathbf{y} = \mathbf{X}\beta + \sum_{r=1}^{R} \mathbf{Z}_r \mathbf{b}_r + \epsilon$$

where $\epsilon \sim N(\mathbf{0}, \mathbf{I})$ independently of $\mathbf{b_r}$. Maximum likelihood estimates are obtained via a Monte Carlo EM algorithm treating the latent variables as the missing data. The E-step is made computationally possible by using Gibbs sampling and the M-step is simplified because of the assumptions of a probit link. An extension of this method to a mixture of binary and continuous responses assuming correlated errors is proposed in this section.

Random effects regression models for ordinal regression have been considered by Harville and Mee (1984), Hedeker and Gibbons (1994) and Tutz and Hennevogl (1996). Tutz and Hennevogl used an EM algorithm treating the random effects and the observed ordinal counts as the complete data. They assumed the thresholds to be unknown and estimated identifiable transformations of them. In contrast, Chan and Kuk's approach would not allow estimation of the thresholds from the complete data if it were to be used for ordinal data. That is why, in the extensions allowing multinomial component considered in Section 5.7 the thresholds are assumed to be known. If they are unknown then the Tutz and Hennevogl algorithm may be extended to handle this case.

A Monte Carlo EM algorithm for multivariate probit model for ordinal data have been considered by Blackwell and Catalano (1999a, 1999b). Catalano (1994) used the GEE approach to fit a model to a bivariate response consisting of an ordinal and a continuous variable. Multivariate tobit analysis has been considered in the econometrics literature (Lee, 1993).

5.2 Model Definition

To develop the correlated probit model for a mixture of a binary and of a continuous response the binary response will be considered as arising from dichotomizing a latent continuous response. Let $\{y_{i1j}\}$ denote the observed continuous measurement and $\{y^*_{i2j}\}$ denote the latent continuous measurement underlying the binary response at the j^{th} occasion for the i^{th} subject, $i = 1, ...n,\ j = 1, ...n_i$. The observed binary variable is then the indicator $y_{i2j} = I\{y^*_{i2j} > 0\}$. The underlying linear mixed model is defined as follows:

$$y_{i1j} = \mathbf{x}^T_{i1j}\boldsymbol{\beta}_1 + \mathbf{z}^T_{i1j}\mathbf{b}_{i1} + \epsilon_{i1j} \tag{5.2}$$

$$y^*_{i2j} = \mathbf{x}^T_{i2j}\boldsymbol{\beta}_2 + \mathbf{z}^T_{i2j}\mathbf{b}_{i2} + \epsilon_{i2j}, \tag{5.3}$$

where $\mathbf{x}_{i1j}$, $\mathbf{x}_{i2j}$, $\mathbf{z}_{i1j}$ and $\mathbf{z}_{i2j}$ are known $p_1 \times 1$, $p_2 \times 1$, $q_1 \times 1$ and $q_2 \times 1$ vectors and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are unknown $p_1 \times 1$ and $p_2 \times 1$ parameter vectors. The random effects and the random errors are assumed to be normally distributed:

$$\mathbf{b}_i = \begin{pmatrix} \mathbf{b}_{i1} \\ \mathbf{b}_{i2} \end{pmatrix} \sim i.i.d.\, \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}) = \mathbf{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}^T_{12} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right) \tag{5.4}$$

$$\boldsymbol{\epsilon}_{ij} = \begin{pmatrix} \epsilon_{i1j} \\ \epsilon_{i2j} \end{pmatrix} \sim i.i.d.\, \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{e}}) = \mathbf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} \sigma^2_{e1} & \sigma_{e12} \\ \sigma_{e12} & \sigma^2_{e2} \end{bmatrix} \right), \tag{5.5}$$

and $\{\mathbf{b}_i\}$ and $\{\boldsymbol{\epsilon}_{ij}\}$ are assumed independent.

The model for the complete data $\{y_{i1j}\}$ and $\{y^*_{i2j}\}$ translates into the following model for the observed data $\{y_{i1j}\}$ and $\{y_{i2j}\}$: Conditional on $\{\mathbf{b}_{i1}\}$ and $\{\mathbf{b}_{i2}\}$

$$\mu_{i1j} = \mathbf{x}^T_{i1j}\boldsymbol{\beta}_1 + \mathbf{z}^T_{i1j}\mathbf{b}_{i1} \tag{5.6}$$

$$\Phi^{-1}(\mu_{i2j}) = \mathbf{x}^T_{i2j}\boldsymbol{\beta}_2 + \mathbf{z}^T_{i2j}\mathbf{b}_{i2}, \tag{5.7}$$

where μ_{i1j} and μ_{i2j} are the conditional means for the two observed variables, and Φ denotes normal cumulative distribution function.

The first equation is the same as in the complete model while the second equation is derived as follows. Conditional on the random effects

$$\mu_{i2j} = P(y^*_{i2j} > 0) = P(\mathbf{x}^T_{i2j}\boldsymbol{\beta}_2 + \mathbf{z}^T_{i2j}\mathbf{b}_{i2} + \epsilon_{i2j} > 0) =$$

$$P(\epsilon_{i2j} > -(\mathbf{x}^T_{i2j}\boldsymbol{\beta}_2 + \mathbf{z}^T_{i2j}\mathbf{b}_{i2})) = \Phi_{e2}(\mathbf{x}^T_{i2j}\boldsymbol{\beta}_2 + \mathbf{z}^T_{i2j}\mathbf{b}_{i2}),$$

where Φ_{e2} is the cumulative distribution function of a $N(0, \sigma^2_{e2})$ random variable. As mentioned in Section 5.1 the variance component σ^2_{e2} corresponding to the Bernoulli response cannot be estimated from the observed data, which indicates only whether the underlying latent variable is positive or negative. For identifiability reasons it is the usual practice in such models to set the variance of the unobserved continuous variable equal to 1, but a different approach will be taken here as discussed later in this chapter.

Certain special cases of the correlated probit model for mixed binary-continuous outcomes are as follows:

1. If only the binary outcome is considered then the model reduces to a multivariate probit model in the terminology of Ochi and Prentice (1984), and of Lessafre and Molenberghs (1991) with equicorrelated variance-covariance structure. The correlation parameter is referred to as 'tetrachoric correlation coefficient'. If an ordinal categorical response is considered instead of a binary response, a multivariate cumulative probit model for the discrete outcome is obtained, and the correlation parameter for the underlying multivariate normal distribution is called a 'polychoric correlation coefficient'.

2. If only the normal response is observed, the model reduces to a general linear model with equicorrelated variance-covariance structure.

3. If $\sigma_{e12} = 0$ then the model is the same as the corresponding bivariate GLMM model in Chapter 3 with a probit link for the Bernoulli response. If, in addition there are only random intercepts for each variable then it is a special case of the Regan

and Catalano model but without modelling the variance-covariance parameters. The latter model can be fitted using an extension of the maximum likelihood method of Ochi and Prentice (1984).

4. The model, as defined, is the same as the original model of Catalano and Ryan (1992) but they rewrote the model in terms of marginal moments and used GEE to obtain estimates. By marginalizing they lost the ability to estimate the model parameters as they were in the original specification and could only test certain hypotheses. If the discrete outcome is ordinal, the model is the same as that of Catalano (1997), who extended the estimation methods of Catalano and Ryan (1992) to a mixture of ordinal and continuous response.

5.3 Maximum Likelihood Estimation

We now show how maximum likelihood estimates can be obtained using a modification of the EM algorithm for the correlated probit model. We first describe the extension of the Chan and Kuk approach using a Monte Carlo EM algorithm, then show how the stochastic approximation approach can be used to speed up the algorithm, and finally discuss standard error approximation.

5.3.1 Monte Carlo EM Algorithm

To describe the Monte Carlo EM algorithm we first define the complete data and show that there are closed form expressions for the complete data maximum likelihood estimates. This leads to a simple M-step in the EM algorithm. Then we derive the conditional expectations needed in each E-step of the algorithm and explain how those are approximated by Monte Carlo sums. At the end we provide a summary of the algorithm and discuss identifiability and convergence of the algorithm.

Complete Data and Complete Data Maximum Likelihood Estimates

The assumption of an underlying continuous variable for the binary response and of an underlying linear mixed model makes the EM algorithm an appealing method for model fitting because the complete data maximum likelihood estimates are easy to compute. The complete data consists of $\mathbf{b}_i$ and $\{\mathbf{y}_{ij}^*\} = \{(y_{i1j}, y_{i2j}^*)^T\}$, $i = 1, ...n$, $j = 1, ...n_i$. Let $\mathbf{x}_{ij} = \begin{pmatrix} \mathbf{x}_{i1j}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{i2j}^T \end{pmatrix}$, $\mathbf{z}_{ij} = \begin{pmatrix} \mathbf{z}_{i1j}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{z}_{i2j}^T \end{pmatrix}$ and $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$. Then the complete data log-likelihood can be written as

$$logL = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i} log|\boldsymbol{\Sigma}_e| - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\mathbf{y}_{ij}^* - \mathbf{x}_{ij}\boldsymbol{\beta} - \mathbf{z}_{ij}\mathbf{b}_i)^T\boldsymbol{\Sigma}_e^{-1}(\mathbf{y}_{ij}^* - \mathbf{x}_{ij}\boldsymbol{\beta} - \mathbf{z}_{ij}\mathbf{b}_i)$$

$$-\frac{1}{2}\sum_{i=1}^{n} log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}\mathbf{b}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{b}_i.$$

Because $\boldsymbol{\Sigma}$ appears only in the second part of the log-likelihood, which is a logarithm of a multivariate normal density, the complete data maximum likelihood estimates of $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{b}_i\mathbf{b}_i^T. \tag{5.8}$$

When the random effects are held fixed the first part of the complete data log-likelihood is also multivariate normal and hence for a fixed $\boldsymbol{\Sigma}_e$, the complete data maximum likelihood estimates for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{n}\sum_{j=1}^{n_i}\mathbf{x}_{ij}^T\boldsymbol{\Sigma}_e^{-1}\mathbf{x}_{ij})^{-1}(\sum_{i=1}^{n}\sum_{j=1}^{n_i}\mathbf{x}_{ij}^T\boldsymbol{\Sigma}_e^{-1}(\mathbf{y}_{ij}^* - \mathbf{z}_{ij}\mathbf{b}_i)). \tag{5.9}$$

Maximizing the complete data profile likelihood, we obtain a closed form expression for the estimate of $\boldsymbol{\Sigma}_e$:

$$\hat{\boldsymbol{\Sigma}}_e = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\mathbf{y}_{ij}^* - \mathbf{x}_{ij}\hat{\boldsymbol{\beta}} - \mathbf{z}_{ij}\mathbf{b}_i)(\mathbf{y}_{ij}^* - \mathbf{x}_{ij}\hat{\boldsymbol{\beta}} - \mathbf{z}_{ij}\mathbf{b}_i)^T. \tag{5.10}$$

As demonstrated by Meng and Rubin (1993), iterating between the last two equations in the EM algorithm will lead to obtaining the true maximum likelihood estimates. This is the so-called ECM (expectation/conditional maximization) algorithm. At each step of the EM algorithm the new estimate of $\boldsymbol{\Sigma}_e$ is computed at the previous value of $\boldsymbol{\beta}$ and then the new value of $\boldsymbol{\Sigma}_e$ is used to update $\boldsymbol{\beta}$.

E-step of the Monte Carlo EM Algorithm

Because we do not observe the random effects and the latent malformation variable, at each E-step we need to compute conditional expectations of the expressions for the maximum likelihood estimates with respect to the observed data evaluated at the current parameter estimates. Following the argument in Chan and Kuk (1997), we now show that all these conditional expectations depend only on two quantities without closed-form expressions: $E(\mathbf{y}_{i2}^*|\mathbf{y}_{i1}, \mathbf{y}_{i2}, \hat{\boldsymbol{\psi}}^{(r)})$ and $Var(\mathbf{y}_{i2}^*|\mathbf{y}_{i1}, \mathbf{y}_{i2}, \hat{\boldsymbol{\psi}}^{(r)})$, where $\mathbf{y}_{i1} = (y_{i11}, ...y_{i1n_i})^T$, $\mathbf{y}_{i2} = (y_{i21}, ...y_{i2n_i})^T$, $\mathbf{y}_{i2}^* = (y_{i21}^*, ...y_{i2n_i}^*)^T$, and $\hat{\boldsymbol{\psi}}^{(r)}$ denotes the parameter vector estimate at the r^{th} step of the EM-algorithm. Let

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_{i1} \\ \vdots \\ \mathbf{x}_{i,n_i} \end{pmatrix}, \mathbf{Z}_i = \begin{pmatrix} \mathbf{z}_{i1} \\ \vdots \\ \mathbf{z}_{i,n_i} \end{pmatrix}, \mathbf{y}_i = \begin{pmatrix} \mathbf{y}_{i1} \\ \vdots \\ \mathbf{y}_{i,n_i} \end{pmatrix}, \mathbf{y}_i^* = \begin{pmatrix} \mathbf{y}_{i1}^* \\ \vdots \\ \mathbf{y}_{i,n_i}^* \end{pmatrix}, \mathbf{b}_i = \begin{pmatrix} \mathbf{b}_{i1} \\ \mathbf{b}_{i2} \end{pmatrix}.$$ From the model definition we obtain the joint distribution of the complete data

$$\begin{pmatrix} \mathbf{y}_i^* \\ \mathbf{b}_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mathbf{X}_i\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{Z}_i\boldsymbol{\Sigma}\mathbf{Z}_i^T + \mathbf{I}_{n_i} \otimes \boldsymbol{\Sigma}_e & \mathbf{Z}_i\boldsymbol{\Sigma} \\ \boldsymbol{\Sigma}\mathbf{Z}_i^T & \boldsymbol{\Sigma} \end{pmatrix} \right]. \tag{5.11}$$

From the properties of the multivariate normal distribution the conditional distribution of the random effect $\mathbf{b}_i$ given the complete response $\mathbf{y}_i^*$ is

$$N(\boldsymbol{\Sigma}\mathbf{Z}_i^T(\mathbf{Z}_i\boldsymbol{\Sigma}\mathbf{Z}_i^T + \mathbf{I}_{n_i} \otimes \boldsymbol{\Sigma}_e)^{-1}(\mathbf{y}_i^* - \mathbf{X}_i\boldsymbol{\beta}), \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{Z}_i^T(\mathbf{Z}_i\boldsymbol{\Sigma}\mathbf{Z}_i^T + \mathbf{I}_{n_i} \otimes \boldsymbol{\Sigma}_e)^{-1}\mathbf{Z}_i\boldsymbol{\Sigma}). \tag{5.12}$$

Let $\mathbf{\Sigma}_{E_i} = \mathbf{I}_{n_i} \otimes \mathbf{\Sigma}_e$ and $\mathbf{\Sigma}_{B_i} = \mathbf{\Sigma}\mathbf{Z}_i^T(\mathbf{Z}_i\mathbf{\Sigma}\mathbf{Z}_i^T + \mathbf{\Sigma}_{E_i})^{-1}$. Then (5.12) can be rewritten as

$$N(\mathbf{\Sigma}_{B_i}(\mathbf{y}_i^* - \mathbf{X}_i\boldsymbol{\beta}), \mathbf{\Sigma} - \mathbf{\Sigma}_{B_i}\mathbf{Z}_i\mathbf{\Sigma}).$$

Therefore,

$$E(\mathbf{b}_i|\mathbf{y}_i) = E[E(\mathbf{b}_i|\mathbf{y}_i^*)|\mathbf{y}_i] = \mathbf{\Sigma}_{B_i}(E(\mathbf{y}_i^*|\mathbf{y}_i) - \mathbf{X}_i\boldsymbol{\beta})$$

and

$$E(\mathbf{b}_i\mathbf{b}_i^T|\mathbf{y}_i) = E[E(\mathbf{b}_i\mathbf{b}_i^T|\mathbf{y}_i^*)|\mathbf{y}_i] =$$

$$= E[Var(\mathbf{b}_i|\mathbf{y}_i^*) + E(\mathbf{b}_i|\mathbf{y}_i^*)E(\mathbf{b}_i^T|\mathbf{y}_i^*)|\mathbf{y}_i] =$$

$$= \mathbf{\Sigma} - \mathbf{\Sigma}_{B_i}\mathbf{Z}_i\mathbf{\Sigma} + \mathbf{\Sigma}_{B_i}E[(\mathbf{y}_i^* - \mathbf{X}_i\boldsymbol{\beta})(\mathbf{y}_i^* - \mathbf{X}_i\boldsymbol{\beta})^T|\mathbf{y}_i]\mathbf{\Sigma}_{B_i}^T.$$

Letting $\hat{\mathbf{V}}_i^{(r)} = E[(\mathbf{y}_i^* - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{(r)})(\mathbf{y}_i^* - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{(r)})^T|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)}]$

$$\hat{\mathbf{\Sigma}}^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{\Sigma}}^{(r)} - \hat{\mathbf{\Sigma}}_{B_i}^{(r)}\mathbf{Z}_i\hat{\mathbf{\Sigma}}^{(r)} + \hat{\mathbf{\Sigma}}_{B_i}^{(r)}\hat{\mathbf{V}}_i^{(r)}\mathbf{\Sigma}_{B_i}^{(r)T}) \tag{5.13}$$

and

$$\hat{\boldsymbol{\beta}}^{(r+1)} = (\sum_{i=1}^{n}\mathbf{X}_i(\hat{\mathbf{\Sigma}}_{E_i}^{(r)})^{-1}\mathbf{X}_i)^{-1}(\sum_{i=1}^{n}\mathbf{X}_i^T(\hat{\mathbf{\Sigma}}_{E_i}^{(r)})^{-1}E(\mathbf{y}_i^* - \mathbf{Z}_i\mathbf{b}_i|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)})) = \tag{5.14}$$

$$(\sum_{i=1}^{n}\mathbf{X}_i(\hat{\mathbf{\Sigma}}_{E_i}^{(r)})^{-1}\mathbf{X}_i)^{-1}(\sum_{i=1}^{n}\mathbf{X}_i^T(\hat{\mathbf{\Sigma}}_{E_i}^{(r)})^{-1}(E(\mathbf{y}_i^*|\mathbf{y}_i) - \mathbf{Z}_i\hat{\mathbf{\Sigma}}_{B_i}^{(r)}(E(\mathbf{y}_i^*|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)}) - \mathbf{X}_i\hat{\boldsymbol{\beta}}^{(r)}))).$$

Finally,

$$\hat{\mathbf{\Sigma}}_e^{(r+1)} = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}E((\mathbf{y}_{ij}^* - \mathbf{x}_{ij}\hat{\boldsymbol{\beta}}^{(r)} - \mathbf{z}_{ij}\mathbf{b}_i)(\mathbf{y}_{ij}^* - \mathbf{x}_{ij}\hat{\boldsymbol{\beta}}^{(r)} - \mathbf{z}_{ij}\mathbf{b}_i)^T|\mathbf{y}_{ij}, \hat{\boldsymbol{\psi}}^{(r)}) =$$

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}E((\mathbf{y}_{ij}^* - \mathbf{x}_{ij}\hat{\boldsymbol{\beta}}^{(r)})(\mathbf{y}_{ij}^* - \mathbf{x}_{ij}\hat{\boldsymbol{\beta}}^{(r)})|\mathbf{y}_{ij}, \hat{\boldsymbol{\psi}}^{(r)})$$

$$-\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}E((\mathbf{y}_{ij}^* - \mathbf{x}_{ij}\hat{\boldsymbol{\beta}}^{(r)})\mathbf{b}_i^T\mathbf{z}_{ij}^T|\mathbf{y}_{ij}, \hat{\boldsymbol{\psi}}^{(r)})$$

$$-\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i} E(\mathbf{z}_{ij}\mathbf{b}_i(\mathbf{y}^*_{ij} - \mathbf{x}_{ij}\hat{\boldsymbol{\beta}}^{(r)})^T|\mathbf{y}_{ij}, \hat{\boldsymbol{\psi}}^{(r)})$$

$$+\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i} E(\mathbf{z}_{ij}\mathbf{b}_i\mathbf{b}_i^T\mathbf{z}_{ij}|\mathbf{y}_{ij}, \hat{\boldsymbol{\psi}}^{(r)}). \tag{5.15}$$

But

$$E(\mathbf{y}^*_{ij}\mathbf{b}_i^T|\mathbf{y}_{ij}) = E[E(\mathbf{y}^*_{ij}\mathbf{b}_i^T|\mathbf{y}^*_{ij})|\mathbf{y}_{ij}] =$$

$$= E\{\mathbf{y}^*_{ij}[E(\mathbf{b}_i^T|\mathbf{y}^*_{ij})|\mathbf{y}_{ij}]\} = E[\mathbf{y}^*_{ij}(\mathbf{y}^*_i - \mathbf{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}_{B_i}^T|\mathbf{y}_{ij}]$$

and hence all conditional expectations depend only on $E(\mathbf{y}^*_i|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)})$ and $Var(\mathbf{y}^*_i|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)})$. Note that $E(y^*_{i1j}|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)}) = y_{i1j}$ for $j = 1, ...n_i$ and therefore we only need to approximate $E(\mathbf{y}^*_{i2}|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)})$ and $Var(\mathbf{y}^*_{i2}|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)})$.

The Gibbs sampler can be used to approximate the integrals above. Notice that

$$f(\mathbf{y}^*_{i2}|\mathbf{y}_i) = f(\mathbf{y}^*_{i2}|\mathbf{y}_{i1}, \mathbf{y}_{i2}) = \frac{f(\mathbf{y}_{i2}|\mathbf{y}_{i1}, \mathbf{y}^*_{i2})f(\mathbf{y}_{i1}, \mathbf{y}^*_{i2})}{f(\mathbf{y}_{i1}, \mathbf{y}_{i2})} =$$

$$= \frac{f(\mathbf{y}_{i2}|\mathbf{y}^*_{i2})f(\mathbf{y}^*_{i2}|\mathbf{y}_{i1})}{f(\mathbf{y}_{i2}|\mathbf{y}_{i1})} =$$

$$= \begin{cases} \frac{f(\mathbf{y}^*_{i2}|\mathbf{y}_{i1})}{c_i} & \text{if } \mathbf{y}_{i2} \in \mathbf{A}_i \\ 0 & \text{otherwise,} \end{cases}$$

where $c_i = P(\mathbf{y}^*_{i2} \in \mathbf{A}_i)$ and $\mathbf{A}_i = \{\mathbf{y}^*_{i2} : y^*_{i2j} > 0 \text{ if } y_{i2j} = 1 \ \& \ y^*_{i2j} < 0 \text{ if } y_{i2j} = 0\}$. Therefore

$$E(\mathbf{y}^*_{i2}|\mathbf{y}_{i1}, \mathbf{y}_{i2}, \hat{\boldsymbol{\psi}}^{(r)}) = \int \mathbf{y}^*_{i2} f(\mathbf{y}^*_{i2}|\mathbf{y}_{i1}, \mathbf{y}_{i2}, \hat{\boldsymbol{\psi}}^{(r)})d\mathbf{y}^*_{i2} =$$

$$= \frac{1}{c_i}\int_{\mathbf{A}_i} \mathbf{y}^*_{i2} f(\mathbf{y}^*_{i2}|\mathbf{y}_{i1}, \hat{\boldsymbol{\psi}}^{(r)})d\mathbf{y}^*_{i2} = \frac{1}{c_i}E[\mathbf{y}^*_{i2} I\{\mathbf{y}^*_{i2} \in \mathbf{A}_i\}|\mathbf{y}_{i1}, \hat{\boldsymbol{\psi}}^{(r)}].$$

A Monte Carlo approximation of $E(\mathbf{y}^*_{i2}|\mathbf{y}_{i1}, \mathbf{y}_{i2}, \hat{\boldsymbol{\psi}}^{(r)})$ is

$$\frac{\frac{1}{m}\sum_{k=1}^{m} \mathbf{y}^{*(k)}_{i2} I\{\mathbf{y}^{*(k)}_{i2} \in \mathbf{A}_i\}}{\frac{1}{m}\sum_{k=1}^{m} I\{\mathbf{y}^{*(k)}_{i2} \in \mathbf{A}_i\}} = \frac{1}{m^*}\sum_{l=1}^{m^*} \mathbf{y}^{*(l)}_{i2}, \tag{5.16}$$

where $\mathbf{y}_{i2}^{*(l)}$ are simulated values from the distribution of $\mathbf{y}_{i2}^{*}|\mathbf{y}_{i1}, \hat{\psi}^{(r)}$. Similarly $Var(\mathbf{y}_{i2}^{*}|\mathbf{y}_{i1}, \mathbf{y}_{i2}, \hat{\psi}^{(r)})$ is approximated by

$$\frac{1}{m^*}\sum_{l=1}^{m^*}\mathbf{y}_{i2}^{*(l)}\mathbf{y}_{i2}^{*(l)^T} - (\frac{1}{m^*}\sum_{l=1}^{m^*}\mathbf{y}_{i2}^{*(l)})(\frac{1}{m^*}\sum_{l=1}^{m^*}\mathbf{y}_{i2}^{*(l)^T}). \tag{5.17}$$

Hence the problem becomes to simulate values from the distribution of $\{\mathbf{y}_{i2}^{*}|\mathbf{y}_{i1}, \hat{\psi}^{(r)}\}$ that fall in $\mathbf{A}_i$. As the conditional distribution of $\mathbf{y}_{i2}^{*}$ given $\mathbf{y}_{i1}$ at each step of the EM algorithm is multivariate normal with known mean and variance, multivariate rejection sampling can be used to generate values. Samples will be generated from the multivariate normal distribution of $\{\mathbf{y}_{i2}^{*}|\mathbf{y}_{i1}, \hat{\psi}^{(r)}\}$ and only those in $\mathbf{A}_i$ will be accepted. But this may be very inefficient and may slow down the algorithm considerably. For example, simulating just 100 values from the initial distribution for one of the 94 subjects in the Ethylene Glycol example took more than 12 minutes. One iteration of the Monte Carlo EM algorithm took more than 1 hour for a simulation sample size of 100. The simulation sample size increases when the algorithm approaches convergence and hence multivariate rejection sampling is not a practically feasible method for simulating values for the MCEM algorithm in this particular example.

A more practical alternative is to use Gibbs sampling. Because the conditional distribution of $\{\mathbf{y}_{i2}^{*}|\mathbf{y}_{i1}, \hat{\psi}^{(r)}\}$ is multivariate normal with known mean and variance, the conditional distribution of $\{y_{i2j}^{*}|\mathbf{y}_{i1}, \mathbf{y}_{i2j}^{*\,-}, \hat{\psi}^{(r)}\}$ is univariate normal with known mean and variance, $j = 1, ...n_i$. Here $\mathbf{y}_{i2j}^{*\,-}$ denotes $\mathbf{y}_{i2}^{*}$ with y_{i2j}^{*} omitted. Also notice that

$$f(y_{i2j}^{*}|\mathbf{y}_{i1}, \mathbf{y}_{i2}, \mathbf{y}_{i2j}^{*\,-}; \hat{\psi}^{(r)}) = f(y_{i2j}^{*}|\mathbf{y}_{i1}, y_{i2j}, \mathbf{y}_{i2j}^{*\,-}; \hat{\psi}^{(r)})$$

that is, the conditional univariate distribution for each latent variable depends only on the binary indicator for that variable and not on the other binary indicators in

the cluster. The proof is as follows:

$$\frac{f(y^*_{i2j}|\mathbf{y}_{i1},\mathbf{y}_{i2},\mathbf{y^*}_{i2j}{}^{-};\hat{\psi}^{(r)})}{f(y^*_{i2j}|\mathbf{y}_{i1},y_{i2j},\mathbf{y^*}_{i2j}{}^{-};\hat{\psi}^{(r)})} =$$

$$= \frac{f(\mathbf{y}_{i1},\mathbf{y}_{i2},\mathbf{y}^*_{i2};\hat{\psi}^{(r)})}{f(\mathbf{y}_{i1},\mathbf{y}_{i2},\mathbf{y^*}_{i2}{}^{-};\hat{\psi}^{(r)})} \Big/ \frac{f(\mathbf{y}_{i1},y_{i2j},\mathbf{y}^*_{i2};\hat{\psi}^{(r)})}{f(\mathbf{y}_{i1},y_{i2j},\mathbf{y^*}_{i2j}{}^{-};\hat{\psi}^{(r)})} =$$

$$= \frac{f(\mathbf{y}_{i1},\mathbf{y}_{i2},\mathbf{y}^*_{i2};\hat{\psi}^{(r)})}{f(\mathbf{y}_{i1},y_{i2j},\mathbf{y}^*_{i2};\hat{\psi}^{(r)})} \Big/ \frac{f(\mathbf{y}_{i1},\mathbf{y}_{i2},\mathbf{y^*}_{i2}{}^{-};\hat{\psi}^{(r)})}{f(\mathbf{y}_{i1},y_{i2j},\mathbf{y^*}_{i2j}{}^{-};\hat{\psi}^{(r)})} =$$

$$= \frac{f(\mathbf{y}^-_{i2}|\mathbf{y}_{i1},y_{i2j},\mathbf{y}^*_{i2};\hat{\psi}^{(r)})}{f(\mathbf{y}^-_{i2}|\mathbf{y}_{i1},y_{i2j},\mathbf{y^*}_{i2j}{}^{-};\hat{\psi}^{(r)})} = 1.$$

Hence, the Gibbs sampler involves simulating values from several truncated univariate normal distributions. Generating values from univariate truncated normal distributions can be carried out in several ways. One possibility is to generate values from the underlying normal distribution and accept only those falling in the area of interest. Another option is based on quantiles of a univariate normal distribution and is now explained in more detail. Suppose that one wants to generate values from the distribution of X which is $N(\mu,\sigma^2)$ truncated above a constant c. Then $Y = \frac{X-\mu}{\sigma}$ is $N(0,1)$ truncated above $\frac{c-\mu}{\sigma}$. If there were no restrictions on Y, $\Phi(Y)$ would be $Uniform(0,1)$, but because Y is truncated, $\Phi(Y)$ can take only values values between $\Phi(\frac{c-\mu}{\sigma})$ and 1. Hence, if random numbers from $Uniform(\frac{c-\mu}{\sigma},1)$ are generated and transformed back to the original scale using the inverse c.d.f. of the standard normal distribution, the resulting variables will have the univariate truncated normal distribution. If $\mathbf{X}$ was truncated below c one must generate values from $Uniform(0,\frac{c-\mu}{\sigma})$ and proceed as above.

To obtain the $(r+1)^{st}$ sample from the distribution of $\{\mathbf{y}^*_{i2}|\mathbf{y}_{i1},y_{i2j},\hat{\psi}^{(r)}\}$ one iteratively generates values from

$f(\mathbf{y}_{i21}^{*(r+1)}|\mathbf{y}_{i1}, y_{i21}, \mathbf{y}_{i21}^{*(r)^-}, \hat{\psi}^{(r)})$

$f(\mathbf{y}_{i22}^{*(r+1)}|\mathbf{y}_{i1}, y_{i22}, \mathbf{y}_{i21}^{*(r+1)}, \mathbf{y}_{i23}^{*(r)}, ..., \mathbf{y}_{i2n_i}^{*(r)}, \hat{\psi}^{(r)})$

...

$f(\mathbf{y}_{i2n_i}^{*(r+1)}|\mathbf{y}_{i1}, y_{i2n_i}, \mathbf{y}_{i2n_i}^{*(r+1)}, \hat{\psi}^{(r)})$.

Monte Carlo EM Algorithm: Summary and Potential Problems

In summary, the Monte Carlo EM algorithm is carried out as follows:

1. Select an initial estimate $\hat{\psi}^{(0)}$ of the parameter vector. Set $r = 1$.

2. Increase r by 1.

E-step: For each subject i, $i = 1, ...n$, generate m^* random samples from the conditional distribution of $\{\mathbf{y}_{i2}^*|\mathbf{y}_{i1}, \mathbf{y}_{i2}; \hat{\psi}^{(r-1)}\}$ using the Gibbs sampler and compute the approximations (5.16) and (5.17).

3. **M-step**: Update the estimate of the parameter vector $\hat{\psi}^{(r)}$ using (5.13), (5.14) and (5.15).

4. Iterate between (2) and (3) until convergence is achieved.

Louis' approximation to the observed information matrix can be used to estimate the standard errors of the parameters (see Chapter 3) but based on dependent random samples obtained using the Gibbs sampler rather than on i.i.d. random samples.

One disadvantage of the Gibbs sampler is that the generated samples are not conditionally independent and therefore extra work is needed to assess convergence of the algorithm. Chan and Kuk (1997) propose to use several independent runs of the algorithm to assess the extent of Monte Carlo variation. Another potential problem with the Monte Carlo EM algorithm arises because σ_{e2}^2 is estimable from the complete data but is not estimable from the observed data. If the usual approach of restricting σ_{e2}^2 to be equal to 1 is adopted, the maximization step becomes more complicated because there is no longer a closed-form expression for $\hat{\mathbf{\Sigma}}_e$. On the other

hand, if σ_{e2}^2 is held unrestricted, then the EM algorithm will likely converge to unique estimates of the fully identifiable ratios $\boldsymbol{\beta}/\sigma_{e2}$, $\boldsymbol{\Sigma}_{12}/\sigma_{e2}$, $\boldsymbol{\Sigma}_{22}/\sigma_{e2}^2$.

Instead of the underlying continuous random variable y_{i2j}^* consider $y_{i2j}^{**} = y_{i2j}^*/\sigma_{e2}$. Then the model in (5.2) and (5.3) can be rewritten as

$$y_{i1j} = \mathbf{x}_{i1j}^T \boldsymbol{\beta}_1 + \mathbf{z}_{i1j}^T \mathbf{b}_{i1} + \epsilon_{i1j}$$

$$y_{i2j}^{**} = \mathbf{x}_{i2j}^T \boldsymbol{\beta}_2^* + \mathbf{z}_{i2j}^T \mathbf{b}_{i2}^* + \epsilon_{i2j}^*,$$

where $\boldsymbol{\beta}_2^* = \boldsymbol{\beta}/\sigma_{e2}$, $\mathbf{b}_{i2}^* = \mathbf{b}_{i2}/\sigma_{e2}$ and $\boldsymbol{\epsilon}_{i2j}^* = \boldsymbol{\epsilon}_{i2j}/\sigma_{e2}$. For the new variables the random effects and the error distributions are as follows:

$$\begin{pmatrix} \mathbf{b}_{i1} \\ \mathbf{b}_{i2}^* \end{pmatrix} \sim i.i.d. \mathbf{N} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}/\sigma_{e2} \\ \boldsymbol{\Sigma}_{12}^T/\sigma_{e2} & \boldsymbol{\Sigma}_{22}/\sigma_{e2}^2 \end{bmatrix} \right)$$

$$\boldsymbol{\epsilon}_{ij} = \begin{pmatrix} \epsilon_{i1j} \\ \epsilon_{i2j}^* \end{pmatrix} \sim i.i.d. \mathbf{N} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{e1}^2 & \rho\sigma_{e1} \\ \rho\sigma_{e1} & 1 \end{bmatrix} \right).$$

As shown by this reparametrization $\boldsymbol{\beta}_2^*/\sigma_{e2}$, $\boldsymbol{\Sigma}_{12}/\sigma_{e2}$ and $\boldsymbol{\Sigma}_{22}/\sigma_{e2}^2$ are identifiable from the observed data.

The resulting EM algorithm when σ_{e2} is held unrestricted can be regarded as a parameter expanded (PX-EM) algorithm (Liu, Rubin and Wu, 1999). In applications when the traditional EM algorithm converges slowly, Liu, Rubin and Wu suggest to expand the complete-data model while preserving the observed-data model and generate an EM using the expanded complete-data model. The idea is that there is extra information in the imputed complete data, which can be used to make the EM algorithm more efficient.

Because of the need to generate a large number of multi-dimensional samples from the conditional distribution of the latent response at each step of the algorithm, the procedure can be very computationally intensive. An alternative fitting procedure is based on an accelerated Monte Carlo EM algorithm proposed by Liao (1999). The

same algorithm has been independently proposed by Lavielle, Delyon, and Moulines (1999), who called it a stochastic approximation EM algorithm (SAEM) because each expectation step of the EM algorithm is replaced by one iteration of a stochastic approximation procedure.

5.3.2 Stochastic Approximation EM Algorithm

The convergence results hold for a complete data log-likelihood of the form

$$ln\ L_{\mathbf{u}}(\psi) = [a(\psi)]^T \mathbf{z}(\mathbf{u}) - b(\mathbf{y}, \psi),$$

where $\mathbf{z}(\mathbf{u})$ is a vector function of the complete data $\mathbf{u}$ and $b(\mathbf{y}, \psi)$ is a function of the observed data $\mathbf{y}$ and the parameter vector ψ but not of the missing data. In the usual Monte Carlo EM algorithm the r^{th} E-step uses the approximations

$$E(\mathbf{z}(\mathbf{u})|\mathbf{y}, \hat{\psi}^{(r)}) \approx \frac{1}{m} \sum_{k=1}^{m} \mathbf{z}(\mathbf{u}^{(k)}),$$

where $\mathbf{u}^{(k)}$, $k = 1, ...m$ are generated values from the conditional distribution of $\{\mathbf{u}|\mathbf{y}, \hat{\psi}^{(r-1)}\}$. Liao proposed to calculate

$$\bar{\mathbf{z}}^{(r)} = (1 - w_r)\bar{\mathbf{z}}^{(r-1)} + w_r \mathbf{z}(\mathbf{u}^{(r)})$$

instead, where $\mathbf{u}^{(r)}$ is only one generated value of the conditional distribution above. The Lavielle et al. (1999) approach is more general because they propose to generate m_r values from $\{\mathbf{u}|\mathbf{y}, \hat{\psi}^{(r-1)}\}$ and take

$$\bar{\mathbf{z}}^{(r)} = (1 - w_r)\bar{\mathbf{z}}^{(r-1)} + w_r[\frac{1}{m_r} \sum_{k=1}^{m_r} \mathbf{z}(\mathbf{u}^{(k)})].$$

Here w_r are chosen weights which must satisfy certain conditions, described by Liao and Lavielle et al. Liao recommends using $w_r = \frac{2}{r+2}$ but other weights may be more

appropriate in certain problems. The modified expectation step is followed by the usual maximization step after an initial 'stabilization period' of length r_0 for $\bar{\mathbf{z}}_r$. This algorithm can be applied to the correlated probit model as follows:

1. Select an initial estimate $\hat{\psi}^{(0)}$ of the parameter vector. Generate r_0 samples from the distribution of $\{\mathbf{y}_{i2}^*|\mathbf{y}_{i1}, \mathbf{y}_{i2}; \hat{\psi}^{(0)}\}$ and compute the approximations

$$\bar{\mathbf{z}}_1^{(r_0)} = \frac{1}{r_0}\sum_{k=1}^{r_0} \mathbf{y}_{i2}^{*\,(k)}$$

$$\bar{\mathbf{z}}_2^{(r_0)} = \frac{1}{r_0}\sum_{k=1}^{r_0} \mathbf{y}_{i2}^{*\,(k)}\mathbf{y}_{i2}^{*\,(k)T}.$$

Set $r = r_0$.

2. Increase r by 1.

E-step: For each subject i, $i = 1, ...n$ generate one random sample $\mathbf{y}_{i2}^{*\,(r)}$ from the conditional distribution of $\mathbf{y}_{i2}^*|\mathbf{y}_{i1}, \mathbf{y}_{i2}; \hat{\psi}^{(r-1)}$ using multivariate rejection sampling and compute the approximations

$$\bar{\mathbf{z}}_1^{(r)} = (1 - w_r)\bar{\mathbf{z}}_1^{(r-1)} + w_r\mathbf{y}_{i2}^{*\,(r)} \tag{5.18}$$

and

$$\bar{\mathbf{z}}_2^{(r)} = (1 - w_r)\bar{\mathbf{z}}_2^{(r-1)} + w_r\mathbf{y}_{i2}^{*\,(r)}\mathbf{y}_{i2}^{*\,(r)T} \tag{5.19}$$

to $E(\mathbf{y}_{i2}^*|\mathbf{y}_{i1}, \mathbf{y}_{i2}; \hat{\psi}^{(r-1)})$ and $E(\mathbf{y}_{i2}^*\mathbf{y}_{i2}^{*T}|\mathbf{y}_{i1}, \mathbf{y}_{i2}; \hat{\psi}^{(r-1)})$ respectively.

3. **M-step:** Update the estimate of the parameter vector $\hat{\psi}^{(r)}$ by substituting $\bar{\mathbf{z}}_1^{(r-1)}$ and $\bar{\mathbf{z}}_2^{(r-1)}$ in (5.13), (5.14) and (5.15).

4. Iterate between (2) and (3) until convergence is achieved.

As in the previous Monte Carlo EM algorithm several runs of the algorithm can be used to assess convergence. The standard errors can be obtained using Louis's method based on additional random samples after the algorithm is stopped but this

will be computationally inefficient and may be subject to similar problems as for the multivariate GLMM. An alternative approach is to apply the same type of stochastic approximation as used for the parameter estimates (Lavielle et al., 1999). We discuss that approach in the next subsection.

One more detail of the algorithm is worth pointing out. Multivariate rejection sampling was suggested to be used to generate values from the needed conditional distributions. This is required to satisfy one of the conditions for convergence as established by Liao (1999) and by Lavielle et al. (1999), namely that conditional on the parameter estimates $\hat{\boldsymbol{\psi}}^{(1)}$, ...$\hat{\boldsymbol{\psi}}^{(r-1)}$, the simulated values $\mathbf{u}^{(1)}$, ... $\mathbf{u}^{(r)}$ are conditionally independent. Lavielle et al. mention that this condition can be relaxed to the case of Markovian dependence, i.e. when conditional on $\hat{\boldsymbol{\psi}}^{(1)}$, ...$\hat{\boldsymbol{\psi}}^{(r-1)}$, $\mathbf{u}^{(r)}$ depends only on $\mathbf{u}^{(r-1)}$ but not on $\mathbf{u}^{(0)}$, ... $\mathbf{u}^{(r-2)}$ but this is still a result that needs to be proved. In the correlated probit model, if the Gibbs sampler can be used instead of multivariate rejection sampling to generate $\mathbf{u}^{(r)}$, the algorithm can achieve convergence much faster. As will be demonstrated in the next section because of the inefficiency of the multivariate rejection sampling Liao's algorithm does not provide an improvement in speed over the Chan and Kuk algorithm for the ethylene glycol data set, unless the Gibbs sampler is used in place of multivariate rejection sampling. The modification to the r^{th} E-step is as follows:

For each subject i, $i = 1, ...n$, generate one random sample $\mathbf{y}_{i2}^{*}$ from the conditional distribution of $\{\mathbf{y}_{i2}^{*}|\mathbf{y}_{i1}, \mathbf{y}_{i2}; \hat{\boldsymbol{\psi}}^{(r-1)}\}$ using the Gibbs sampler (as described in Chan and Kuk's algorithm) with starting values $\mathbf{y}_{i2}^{*(r-1)}$ and compute the approximations (5.18) and (5.19).

5.3.3 Standard Error Approximation

As outlined in Chapter 3, the observed data information matrix can be represented as follows:

$$-\frac{\partial^2 l(\mathbf{y},\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T} = -E[\frac{\partial^2 lnL_u(\mathbf{b},\mathbf{y}^*,\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T}|\mathbf{y}] - Var[\frac{\partial lnL_u(\mathbf{b},\mathbf{y}^*,\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}|\mathbf{y}] =$$

$$-E[\frac{\partial^2 lnL_u(\mathbf{b},\mathbf{y}^*,\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T} + \frac{\partial lnL_u(\mathbf{b},\mathbf{y}^*,\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\frac{\partial lnL_u(\mathbf{b},\mathbf{y}^*,\boldsymbol{\psi})}{\partial\boldsymbol{\psi}^T}|\mathbf{y}]$$

$$+E[\frac{\partial lnL_u(\mathbf{b},\mathbf{y}^*,\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}|\mathbf{y}]E[\frac{\partial lnL_u(\mathbf{b},\mathbf{y}^*,\boldsymbol{\psi})}{\partial\boldsymbol{\psi}^T}|\mathbf{y}] \doteq \mathbf{G} + \boldsymbol{\Delta}\boldsymbol{\Delta}^T.$$

By simulating values from the conditional distribution of $\{(\mathbf{b},\mathbf{y}^*)|\mathbf{y}\}$, both G and $\boldsymbol{\Delta}$ above can be approximated at each step of the algorithm:

$$\mathbf{G}^{(r)} = (1-w_r)\mathbf{G}^{(r-1)} + w_r\mathbf{N}_r,$$

where $\mathbf{N}_r = \frac{\partial^2 lnL_u(\mathbf{b}^{(r)},\mathbf{y}^{*(r)},\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T} + \frac{\partial lnL_u(\mathbf{b}^{(r)},\mathbf{y}^{*(r)},\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\frac{\partial lnL_u(\mathbf{b}^{(r)},\mathbf{y}^{*(r)},\boldsymbol{\psi})}{\partial\boldsymbol{\psi}^T}$ and

$$\boldsymbol{\Delta}_r = (1-w_r)\boldsymbol{\Delta}_{r-1} + w_r\frac{\partial lnL_u(\mathbf{b}^{(r)},\mathbf{y}^{*(r)},\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}.$$

These are the approximations for $r > r_0$. For $r = r_0$

$$\mathbf{G}_r = \frac{1}{r_0}\sum_{k=1}^{r_0}\{\frac{\partial^2 lnL_u(\mathbf{b}^{(k)},\mathbf{y}^{*(k)},\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^T} + \frac{\partial lnL_u(\mathbf{b}^{(k)},\mathbf{y}^{*(k)},\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\frac{\partial lnL_u(\mathbf{b}^{(k)},\mathbf{y}^{*(k)},\boldsymbol{\psi})}{\partial\boldsymbol{\psi}^T}\}$$

$$\boldsymbol{\Delta}_r = \frac{1}{r_0}\sum_{k=1}^{r_0}\frac{\partial lnL_u(\mathbf{b}^{(k)},\mathbf{y}^{*(k)},\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}.$$

It should be noted that it was not necessary to generate values from the random effects distribution for the parameter estimation but for the standard errors estimation this

can not be avoided. Fortunately not much extra effort is required and the modification can be incorporated in the algorithm easily. Notice that

$$f(\mathbf{y}^*, \mathbf{b}|\mathbf{y}) = f(\mathbf{y}^*|\mathbf{y}) f(\mathbf{b}|\mathbf{y}, \mathbf{y}^*) = f(\mathbf{y}^*|\mathbf{y}) f(\mathbf{b}|\mathbf{y}^*),$$

so one can first simulate $\mathbf{y}_2^*$ as outlined for the parameter estimation and then simulate $\mathbf{b}$ from $\mathbf{b}|\mathbf{y}^*$, which is multivariate normal.

5.4 Application

The ethylene glycol data can be analyzed using the method outlined in the previous sections if we assume an underlying latent malformation variable. The model will then be as follows

y_{i1j} - fetal weight of j^{th} live fetus in i^{th} litter.

y_{i2j}^* - latent malformation of j^{th} live fetus in i^{th} litter.

$y_{i2j} = I\{y_{i2j}^* > 0\}$ - observed malformation status of j^{th} live fetus in i^{th} litter.

$$y_{i1j} = \beta_{10} + d_i * \beta_{11} + b_{i1} + \epsilon_{i1j},$$

$$y_{i2j}^* = \beta_{20}^* + d_i * \beta_{21}^* + b_{i2} + \epsilon_{i2j},$$

$$\mathbf{b}_i = (b_{i1}, b_{i2})^T \sim N(\mathbf{0}, \mathbf{\Sigma}),\ \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix},$$

$$\boldsymbol{\epsilon}_{ij} = \begin{pmatrix} \epsilon_{i1j} \\ \epsilon_{i2j} \end{pmatrix} \sim i.i.d.\, \mathbf{N}(\mathbf{0}, \mathbf{\Sigma_e}) = \mathbf{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{e1}^2 & \sigma_{e12} \\ \sigma_{e12} & \sigma_{e2}^2 \end{bmatrix} \right).$$

As seen from the last expression this model is more general than a bivariate GLMM for a binary and a continuous outcome because it allows different correlation between malformation and fetal weight within fetus and between two different fetuses within litter. When $\sigma_{e12} = 0$, this model reduces to the model in Chapter 3 but with a probit

Table 5.1. Maximum likelihood estimates for the ethylene glycol example using the Chan and Kuk method, and the two versions of Liao's method based on multivariate rejection sampling and the Gibbs sampler, respectively.

Parameter	Chan Kuk	Liao - MRS	Liao - Gibbs
β_{10}	0.952	0.952	0.952
β_{11}	-0.087	-0.087	-0.087
β_{20}	-2.377	-2.388	-2.369
β_{21}	0.969	0.967	0.966
σ_{b1}	0.086	0.086	0.086
σ_{b2}	0.824	0.844	0.822
ρ_b	-0.607	-0.633	-0.608
σ_{e1}	0.075	0.075	0.075
ρ_e	-0.202	-0.203	-0.201

instead of a logit link for malformation. Testing $\sigma_{e12} = 0$ is then equivalent to testing conditional independence between the two outcomes.

The identifiable parameters in the above specification will be β_{10}, β_{11}, $\beta_{20} = \frac{\beta^*_{20}}{\sigma_{e2}}$, $\beta_{21} = \frac{\beta^*_{21}}{\sigma_{e2}}$, $\sigma_{b1} = \sigma_1$, $\sigma_{b2} = \frac{\sigma_2}{\sigma_{e2}}$, $\rho_b = \frac{\sigma_{12}}{\sigma_1\sigma_2}$, σ_{e1} and $\rho_e = \frac{\sigma_{e12}}{\sigma_{e2}\sigma_{e1}}$. Table 5.1 contains the estimates for the above model obtained using Chan and Kuk's algorithm and the two versions of Liao's algorithm. Initial estimates for $\boldsymbol{\beta}$ were the regression parameter estimates from the fixed effects models for the two responses and $\hat{\sigma}^{(0)}_{e1}$ was set equal to the estimated standard deviation from the linear mixed model for fetal weight. For the remaining variance components we used arbitrary values: $\hat{\boldsymbol{\Sigma}}^{(0)} = \mathbf{I}_2$, $\hat{\sigma}^{(0)}_{e2} = 1$ and $\hat{\sigma}^{(0)}_{e12} = 0$.

A comparison of the times until convergence for the parameter estimates is provided in Figures 5.1-5.4. The times are actual times (not CPU times) and are given in minutes. For comparison purposes and to remove extraneous sources of variability two algorithms were run on the same Sun Ultra 10 Workstation simultaneously. There were no other big jobs running at the same time. Initially, Chan and Kuk's algorithm and Liao's algorithm using the Gibbs sampler were run simultaneously using the same number of simulated samples (10,000 in each case). This translated into

10,000 iterations for Liao's algorithm and 20 iterations for Chan and Kuk's algorithm with a simulation sample size for the Gibbs sampler of 500 (a burn-in of 100) at each iteration. This number of iterations, however, was not sufficient for convergence of Chan and Kuk's algorithm and hence the algorithms were rerun for 23 hours (corresponding to 137 iterations for Chan and Kuk's algorithm). Liao's algorithm using multivariate rejection sampling was run for 1000 iterations simultaneously with Chan and Kuk's algorithm. 1000 iterations took 60 hours and some of the parameters had not converged yet (see the estimates for σ_{b2}/σ_{e2} and ρ_b in Figure 5.3). As can be easily seen from the graphs Liao's algorithm using the Gibbs sampler is much faster than Chan and Kuk's and Liao's algorithm using multivariate rejection sampling. A more detailed look at Liao's Gibbs sampler (see Figures 5.5-5.8) shows that the estimates appear to have converged (or nearly converged) after only about 500 iterations and in less than half an hour. The other two algorithms take several hours. All three algorithms proved to have convergence problems when the initial estimates were chosen far away from the true values.

Note that we used a sample size of 500 with a burn-in of 100 for the Gibbs sampler at each iteration of Chan and Kuk's algorithm. This is very inefficient especially when the estimates are still far away from the true maximum likelihood estimates but it is unclear how to adaptively increase the sample size. Potentially, Chan and Kuk's algorithm can be accelerated significantly but it is unlikely that it will outperform Liao's algorithm using the Gibbs sampler.

Standard error computations were incorporated in the Liao's Gibbs sampler algorithm. The results after 10,000 iterations for a quadratic trend in dose for fetal weight are given in Table 5.2. The quadratic trend was not significant ($p - value = 0.22$) so the final model assumed only a linear relationship between dose and both responses. Table 5.3 contains the results from the model fit of the final model and also of the fit of two additional reduced models, which will be discussed later in this section.

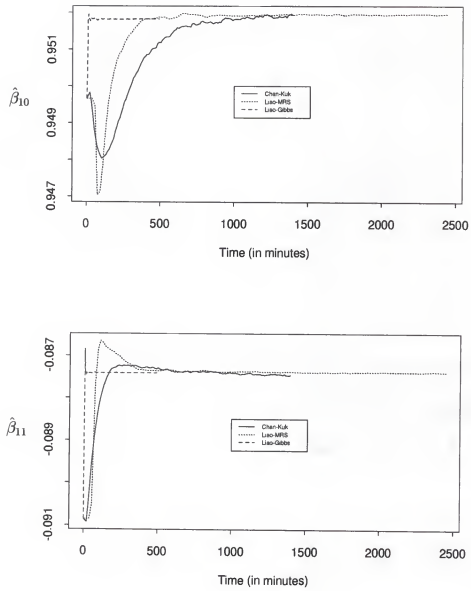

Figure 5.1 Convergence of the intercept estimates ($\hat{\beta}_{10}$) and the slope estimates ($\hat{\beta}_{11}$) for fetal weight in the ethylene glycol example.

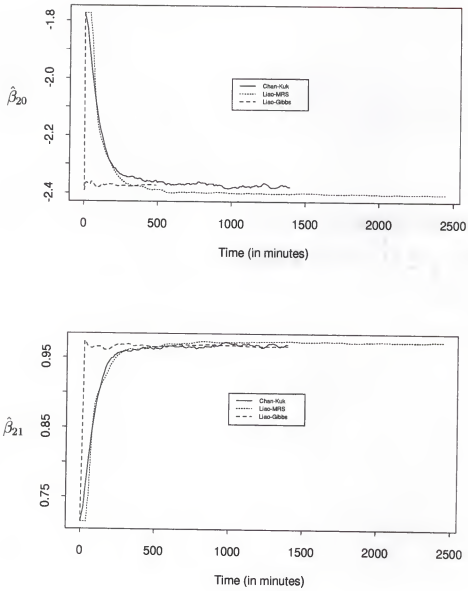

Figure 5.2 Convergence of the intercept estimates ($\hat{\beta}_{20}$) and the slope estimates ($\hat{\beta}_{21}$) for malformation in the ethylene glycol example.

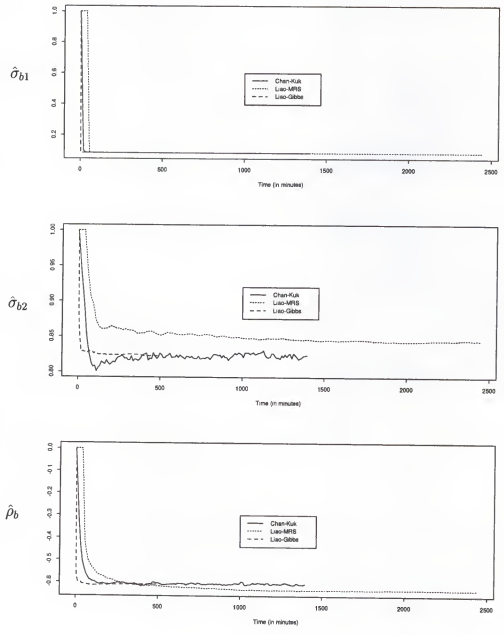

Figure 5.3 Convergence of the variance component estimates for the random effects in the ethylene glycol example.

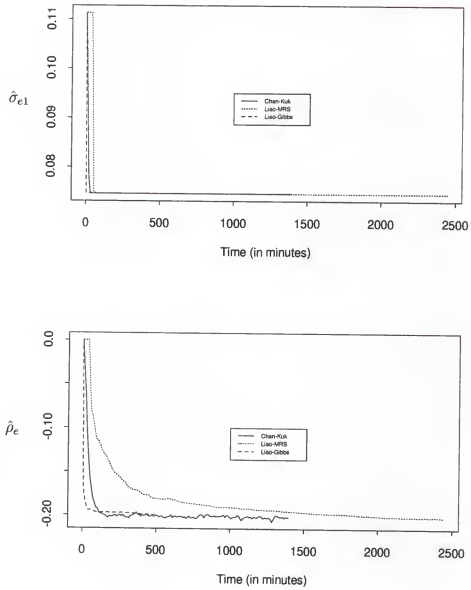

Figure 5.4 Convergence of the variance component estimates for the random errors in the ethylene glycol example.

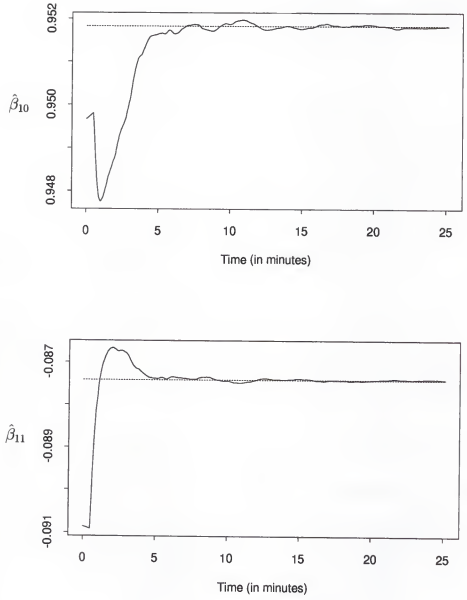

Figure 5.5 Convergence of the intercept and slope estimates for fetal weight in the ethylene glycol example. The estimates are from Liao's method using the Gibbs sampler. The horizontal lines represent the final values of the estimates.

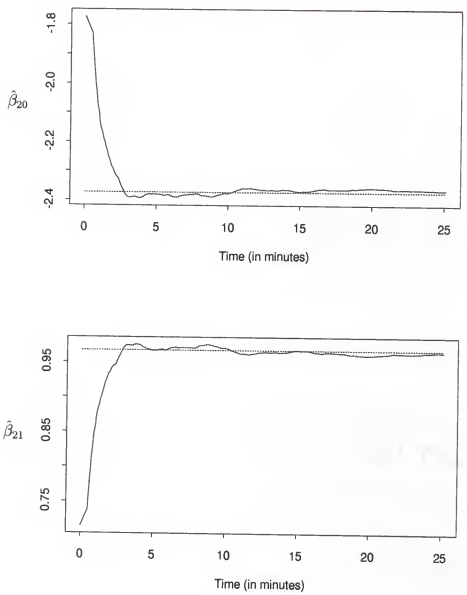

Figure 5.6 Convergence of the intercept and slope estimates for malformation in the ethylene glycol example. The estimates are from Liao's method using the Gibbs sampler. The horizontal lines represent the final values of the estimates.

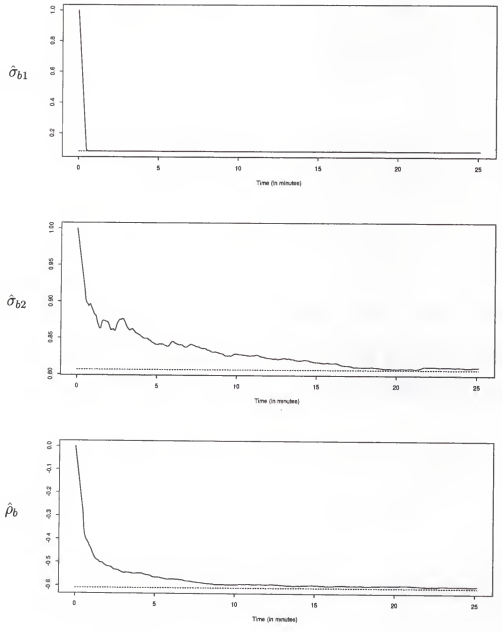

Figure 5.7 Convergence of the variance component estimates for the random effects in the ethylene glycol example. The estimates are from Liao's method using the Gibbs sampler. The horizontal lines represent the final values of the estimates.

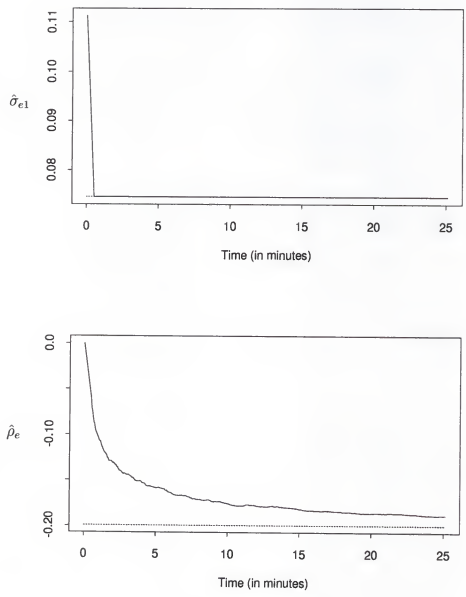

Figure 5.8 Convergence of the variance component estimates for the random errors in the ethylene glycol example. The estimates are from Liao's method using the Gibbs sampler. The horizontal lines represent the final values of the estimates.

Table 5.2. Maximum likelihood estimates from a correlated probit model with a quadratic trend for fetal weight in the ethylene glycol example.

Parameter	Estimate	Standard error
β_{10}	0.963	0.017
β_{11}	-0.120	0.028
β_{12}	0.011	0.009
β_{20}	-2.363	0.214
β_{21}	0.957	0.110
σ_{b1}	0.083	0.007
σ_{b2}	0.836	0.107
ρ_b	-0.608	0.101
σ_{e1}	0.075	0.002
ρ_e	-0.211	0.055

Table 5.3. Maximum likelihood estimates with linear dose effects for both variables in the ethylene glycol example.

Parameter	Full model		Reduced model 1		Reduced model 2	
	Est.	SE	Est.	SE	Est.	SE
β_{10}	0.952	0.014	0.952	0.014	0.952	0.014
β_{11}	-0.087	0.008	-0.087	0.008	-0.087	0.008
β_{20}	-2.396	0.216	-2.401	0.216	-2.416	0.217
β_{21}	0.971	0.110	0.972	0.110	0.988	0.112
σ_{b1}	0.086	0.007	0.086	0.007	0.086	0.007
σ_{b2}	0.837	0.106	0.839	0.107	0.873	0.113
ρ_b	-0.640	0.091	-0.664	0.091	–	–
σ_{e1}	0.075	0.002	0.075	0.002	0.075	0.002
ρ_e	-0.211	0.055	–	–	–	–

As expected the standard error estimates converged more slowly than the parameter estimates (Figures 5.9 and 5.10).

As pointed out by Lavielle et al. (1999) $\mathbf{\Delta}^{(r)} \to 0$ for $r \to \infty$ and therefore the limiting value of $\mathbf{G}^{(r)}$ can be used to assess the variability of the estimators. In practice, however, the algorithm may be stopped before reaching the maximum-likelihood estimator and it seems more reasonable to use both $\mathbf{G}^{(r)}$ and $\mathbf{\Delta}^{(r)}$ to approximate the variances. In fact, in this particular example the two approximations lead to significantly different results (Table 5.4). The standard errors in the column SE_2 are based only on $\mathbf{G}^{(r)}$, while those in the column SE_3 are based on both $\mathbf{G}^{(r)}$ and $\mathbf{\Delta}^{(r)}$.

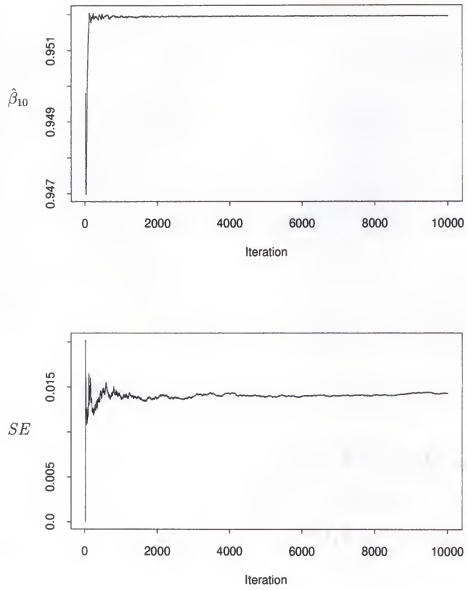

Figure 5.9 Convergence of the intercept estimate and its estimated standard error for fetal weight in the ethylene glycol example.

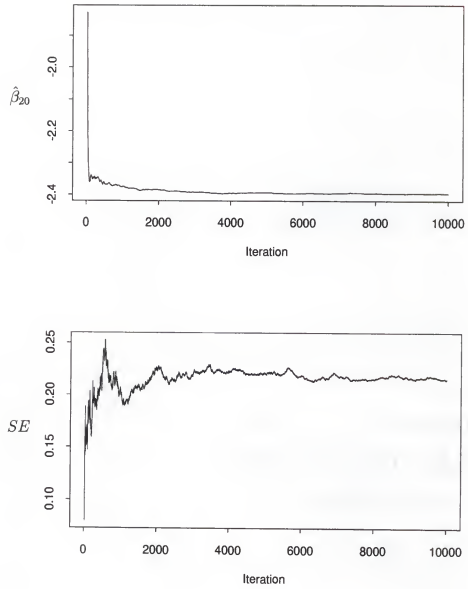

Figure 5.10 Convergence of the intercept estimate and its estimated standard error for malformation in the ethylene glycol example.

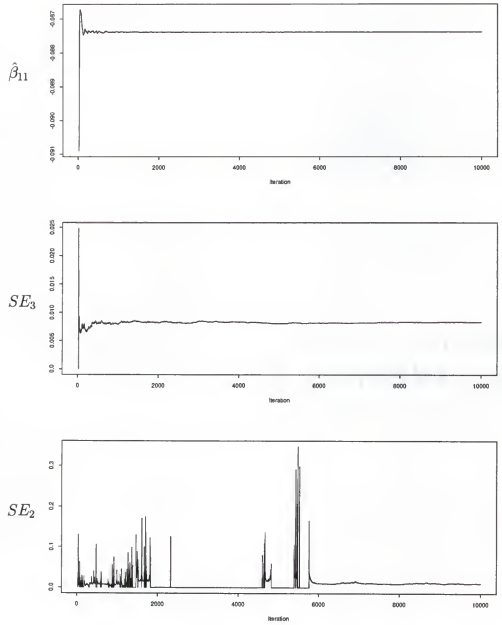

Figure 5.11 Convergence of the slope estimate and its estimated standard errors for fetal weight. SE_2 and SE_3 denote the standard errors corresponding to two and three-term approximations of the observed information matrix respectively.

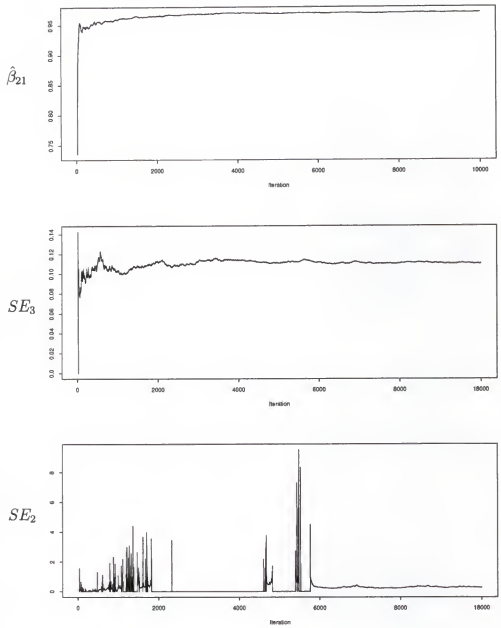

Figure 5.12 Convergence of the slope estimate and its estimated standard errors for malformation. SE_2 and SE_3 denote the standard errors corresponding to two and three-term approximations of the observed information matrix respectively.

Table 5.4. Standard error estimates using two and three-term approximations of the observed information matrix in the ethylene glycol example.

Parameter	SE_2	SE_3
β_{10}	0.038	0.014
β_{11}	0.011	0.008
β_{20}	0.702	0.215
β_{21}	0.270	0.110
σ_{b1}	0.008	0.007
σ_{b2}	0.300	0.107
σ_b	0.223	0.092
σ_{e1}	0.001	0.002
ρ_e	0.000	0.056

The latter ones are the same to within 2 decimal points as the errors obtained using Gaussian quadrature for the corresponding multivariate GLMM. Those based only on $\mathbf{G}^{(r)}$, however, are either much bigger (sometimes more than three times bigger), or equal to zero. Also, Figures 5.11 and 5.12 show typical behaviour of the standard errors for the ethylene glycol data example. SE_3 are shown in the second graphs and SE_2 are shown in the third graphs. Notice that SE_2 are the correct standard errors only asymptotically and there is no guarantee that they will be even close to the right quantities when the algorithm is stopped. The problem most likely is numerical instability, which is compensated for in SE_3 but not in SE_2. Hence, it is strongly suggested to use SE_3.

As mentioned above, Table 5.3 contains the results from three model fits using Liao's Gibbs sampler. In the first reduced model $\rho_e = 0$ by definition; i.e. this is a multivariate GLMM with probit link for malformation, and in the second one both $\rho_e = 0$ and $\rho_b = 0$; i.e. this corresponds to two separate GLMM's for the two response variables. It is clear from the table that there is a significant (but weak) intra-fetus correlation between malformation and fetal weight ($\frac{\hat{\rho}_e}{\widehat{SE}(\hat{\rho}_e)} = \frac{-0.211}{0.055} = -3.84$, $p - value < 0.0001$) but the regression parameter estimates are almost identical to those from the fit of the multivariate GLMM. Even more surprising, the standard

error estimates from all three models are identical up to two significant digits after the decimal point. This is somewhat disappointing because it is expected that when the responses are fitted together and also a better correlation structure is assumed there will be efficiency gains in estimating the parameters. The next section is devoted to this issue.

5.5 Simulation Study

It was of interest to investigate possible efficiency gains in fitting the correlated probit model instead of fitting the corresponding multivariate GLMM, and instead of fitting separate univariate GLMM's for the individual response variables. A simulation study was designed as follows. The structure of the data was assumed to be the same as in the ethylene glycol example. The parameter values (except for the error correlation parameter ρ_e) were set equal to the final estimates from Liao's Gibbs sampler. There were six settings corresponding to one large and two small data sets, and to strong and weak intrafetus correlation ($\rho_e = -0.804$ and $\rho_e = -0.201$). The large data set had numbers of clusters and observations exactly as in the ethylene glycol example (94 clusters and an average of about 11 observations per cluster), both small data sets had 6 observations per cluster but one had 32 clusters and the other one had 24 clusters. The simulation with 32 clusters was added after the other simulations were finished.

A total of 50 samples were generated at each of the six settings and three models were fitted to each of those samples: a corrrelated probit model, referred to as the full model (FM); a multivariate GLMM, referred to as reduced model 1 (RM1), and two separate GLMM's, referred to as reduced model 2 (RM2). All models were fit using Liao's Gibbs sampler but in RM1 ρ_e was assumed to be equal to 0, and in RM2 both ρ_e and ρ_b were assumed to be equal to 0. The algorithms for the larger data set were run for 5000 iterations while those for the smaller data sets were run for 10000

iterations. The standard errors were computed using the stochastic approximation approach. The initial estimates for the regression parameters were the estimates from the fixed effects fits. The sample standard deviation was used as an initial estimate for σ_{e1} and some "reasonable" values were used for the other variance components $\hat{\sigma}_{b1}^{(0)} = \hat{\sigma}_{b2}^{(0)} = 1$, $\hat{\rho}_b^{(0)} = -0.5$ and $\hat{\rho}_e^{(0)} = -0.5\hat{\sigma}_{e1}^{(0)}$. In the reduced models the same initial parameter values were used except for those parameters that were not included in the models.

The simulation programs for the larger data sets ran for about 25 days each on Sun Ultra 10 Workstation with 128 MB of RAM. The simulation programs for the smaller data sets ran for about 10 days. The results are summarized in Tables 5.5 - 5.16. The data set with 32 clusters is referred to as the medium data set. For each setting there are two tables: one gives the average parameter estimates and the average standard errors, and the other one gives the standard deviations for the parameter and standard error estimates. By comparing the average of the standard error estimates and the standard deviations for the parameters, Monte Carlo error can be judged.

For the large data set the sample standard deviations of the estimated parameters are very similar to the corresponding means of the estimated standard errors, indicating that the Monte Carlo error is small (see Tables 5.5 - 5.8). This however, is not true for the small simulation sample size for the estimates of some of the parameters (β_{20}, β_{21}, σ_{b2}, ρ_b and ρ_e) in the full model and to a lesser extent in the reduced models. For example, the standard deviations of $\hat{\beta}_{20}$ and $\hat{\beta}_{21}$ in the full model are larger that the mean standard errors. (0.442 and 0.213 as compared to 0.323 and 0.175 in Tables 5.11 and 5.12). Similar observations can be made for the other small sample settings. It is not surprising that Monte Carlo error is increased when the observations per cluster and the number of clusters are reduced but requires caution in interpreting the simulation study results.

When the sample size is large and the correlation is weak the average standard errors for all parameters are very similar and the largest difference occurs for β_{20} (0.211 compared to 0.216 in Table 5.5). There is hardly any gain in efficiency in fitting the responses together rather than fitting them separately although both ρ_b and ρ_e are significantly different from zero. When the correlation ρ_e is strong, there is very slight gain in efficiency for the regression parameters for the Bernoulli response (0.209 in the FM compared to 0.221 for RM1 and to 0.223 for RM2 for β_{20}, and 0.107 in the FM compared to 0.113 in RM1 and RM2 for β_{21} in Table 5.7). For small sample size and strong intrafetus correlation (ρ_e) the efficiency gains are much more pronounced: the average standard error estimate for β_{21} goes up from 0.175 to 0.255 between FM and RM2 (Table 5.11). Slightly smaller difference is observed in the standard deviations (Table 5.12). Although such increase in the standard errors will not lead to a different conclusion about the significance of the effects, it is important when setting up confidence intervals to estimate strength of effects. Some efficiency gains are also observed in the small sample case when the intrafetus correlation is weak, although this may at least partially be due to Monte Carlo error (Tables 5.9 and 5.10). Notice that in that case the intrafetus correlation is not significant ($p - value = 0.24$). The simulation results for the medium data set were very similar to those for the small data set and will not be discussed in more detail here.

In summary, it appears that there are noticeable efficiency gains in parameter estimation only for small data sets and strong correlations between the responses within cluster. The efficiency gains are the largest for the binary response regression parameters. The cases considered are few and so recommendations can be only tentative. However, it seems that unless the provided sample is small and there is evidence of strong intracluster correlations between the response variables, it may not be worth the extra computational effort to fit the responses jointly over fitting them separately. Joint fitting may still be necessary in the cases when multivariate questions must be

answered and to obtain the true maximum likelihood estimates when the correlated probit model is the true underlying model, but the efficiency gains may not be great.

5.6 Identifiability Issue

For reasons of computational simplicity the variance component σ_2^2 was left unrestricted in the EM algorithm decribed in Section 5.2. It was hypothesized that the EM algorithm will converge to unique estimates of the identifiable parameters. We now address this issue in greater detail.

Let us consider a simpler case where there is a closed form expression for the maximum likelihood estimates. Suppose that we have n i.i.d. Bernoulli random variables $y_i \sim Be(\pi)$, $i = 1, ...n$, and assume that they arise from dichotomizations of n i.i.d. unobserved Normal random variables $y_i^* \sim N(\mu, \sigma)$, that is $y_i = I\{y_i^* > 0\}$. Then $\pi = \Phi(\frac{\mu}{\sigma})$ as demonstrated earlier in this chapter. The maximum likelihood estimate of π is $\frac{n_1}{n}$, where n_1 is the number of ones in the observed sample, and hence the maximum likelihood estimate of $\frac{\mu}{\sigma}$ is $\Phi^{-1}(\frac{n_1}{n})$. μ and σ are not individually estimable from the observed data, but are estimable from the complete data y_i^*, $i = 1, ...n$. For this simple case we can derive expressions for the estimates of μ and σ at the $(r+1)^{st}$ step of the EM algorithm in terms of the previous estimates at the r^{th} step and show that the EM algorithm converges to the true maximum lieklihood estimate of $\frac{\mu}{\sigma}$. Moreover, the estimates of μ and σ converge to some values μ^* and σ^* such that the ratio $\frac{\mu^*}{\sigma^*}$ is the maximum likelihood estimate.

The complete data log-likelihood is

$$l_c = -\frac{n}{2} log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i^* - \mu)^2.$$

Table 5.5. Results from simulation study: average estimates and average estimated standard errors for large sample and weak correlation.

Parameter	True	Full model		Reduced model 1		Reduced model 2	
	value	Est.	SE	Est.	SE	Est.	SE
β_{10}	0.952	0.955	0.014	0.955	0.014	0.955	0.014
β_{11}	-0.087	-0.089	0.008	-0.089	0.008	-0.089	0.008
β_{20}	-2.369	-2.471	0.211	-2.462	0.209	-2.466	0.216
β_{21}	0.966	1.016	0.108	1.012	0.106	1.014	0.109
σ_{b1}	0.086	0.086	0.007	0.086	0.007	0.086	0.007
σ_{b2}	0.822	0.804	0.114	0.799	0.114	0.796	0.116
ρ_b	-0.608	-0.620	0.098	-0.642	0.098	–	–
$\sigma_{\epsilon 1}$	0.075	0.075	0.002	0.075	0.002	0.075	0.002
ρ_ϵ	-0.201	-0.195	0.059	–	–	–	–

Table 5.6. Results from simulation study: standard deviations of estimates and of estimated standard errors for large sample and weak correlation

Parameter	Full model		Reduced model 1		Reduced model 2	
	Est.	SE	Est.	SE	Est	SE
β_{10}	0.014	0.00113	0.014	0.00112	0.014	0.00108
β_{11}	0.008	0.00068	0.008	0.00065	0.008	0.00063
β_{20}	0.197	0.02617	0.194	0.02235	0.202	0.02639
β_{21}	0.115	0.01431	0.114	0.01049	0.119	0.01199
σ_{b1}	0.007	0.00047	0.007	0.00047	0.007	0.00047
σ_{b2}	0.102	0.01228	0.100	0.01163	0.102	0.01226
ρ_b	0.098	0.01687	0.097	0.01692	–	–
$\sigma_{\epsilon 1}$	0.002	0.00004	0.002	0.00004	0.002	0.00004
ρ_ϵ	0.059	0.00277	–	–	–	–

Table 5.7. Results from simulation study: average estimates and average estimated standard errors for large sample and strong correlation.

Parameter	True	Full model		Reduced model 1		Reduced model 2	
	value	Est.	SE	Est.	SE	Est.	SE
β_{10}	0.952	0.951	0.014	0.951	0.014	0.951	0.014
β_{11}	-0.087	-0.088	0.008	-0.088	0.008	-0.088	0.008
β_{20}	-2.369	-2.446	0.209	-2.499	0.221	-2.499	0.223
β_{21}	0.966	1.005	0.107	1.027	0.113	1.028	0.113
σ_{b1}	0.086	0.084	0.007	0.085	0.007	0.084	0.007
σ_{b2}	0.822	0.808	0.115	0.847	0.120	0.844	0.123
ρ_b	-0.608	-0.615	0.092	-0.686	0.089	–	–
σ_{e1}	0.075	0.075	0.002	0.075	0.002	0.075	0.002
ρ_e	-0.801	-0.766	0.031	–	–	–	–

Table 5.8. Results from simulation study: standard deviations of estimates and of estimated standard errors for large sample and strong correlation.

Parameter	Full model		Reduced model 1		Reduced model 2	
	Est.	SE	Est.	SE	Est	SE
β_{10}	0.015	0.00094	0.015	0.00092	0.015	0.00090
β_{11}	0.008	0.00054	0.008	0.00051	0.008	0.00057
β_{20}	0.207	0.03556	0.218	0.03760	0.229	0.03145
β_{21}	0.104	0.01602	0.112	0.01770	0.116	0.01510
σ_{b1}	0.005	0.00037	0.005	0.00037	0.005	0.00037
σ_{b2}	0.114	0.01579	0.122	0.01705	0.128	0.01602
ρ_b	0.095	0.01476	0.095	0.01492	–	–
σ_{e1}	0.002	0.00004	0.002	0.00004	0.002	0.00004
ρ_e	0.027	0.00330	–	–	–	–

Table 5.9. Results from simulation study: average estimates and average estimated standard errors for small sample and weak correlation.

Parameter	True	Full model		Reduced model 1		Reduced model 2	
	value	Est.	SE	Est.	SE	Est.	SE
β_{10}	0.952	0.950	0.028	0.950	0.028	0.950	0.028
β_{11}	-0.087	-0.085	0.017	-0.085	0.016	-0.085	0.016
β_{20}	-2.369	-2.610	0.489	-2.567	0.560	-2.618	0.605
β_{21}	0.966	1.054	0.243	1.038	0.267	1.060	0.282
σ_{b1}	0.086	0.083	0.014	0.083	0.013	0.083	0.014
σ_{b2}	0.822	0.787	0.225	0.772	0.262	0.761	0.320
ρ_b	-0.608	-0.646	0.177	-0.725	0.141	–	–
σ_{e1}	0.075	0.075	0.005	0.075	0.005	0.075	0.005
ρ_e	-0.201	-0.188	0.159	–	–	–	–

Table 5.10. Results from simulation study: standard deviations of estimates and of estimated standard errors for small sample and weak correlation.

Parameter	Full model		Reduced model 1		Reduced model 2	
	Est.	SE	Est.	SE	Est	SE
β_{10}	0.027	0.004	0.027	0.004	0.027	0.004
β_{11}	0.014	0.002	0.014	0.002	0.014	0.002
β_{20}	0.734	0.160	0.648	0.582	0.802	0.393
β_{21}	0.300	0.069	0.273	0.247	0.339	0.168
σ_{b1}	0.012	0.002	0.012	0.006	0.012	0.002
σ_{b2}	0.345	0.098	0.340	0.507	0.380	0.166
ρ_b	0.315	0.151	0.202	0.141	–	–
σ_{e1}	0.004	0.0003	0.004	0.0005	0.004	0.0003
ρ_e	0.149	0.023	–	–	–	–

Table 5.11. Results from simulation study: average estimates and average estimated standard errors for small sample and strong correlation.

Parameter	True	Full model		Reduced model 1		Reduced model 2	
	value	Est.	SE	Est.	SE	Est.	SE
β_{10}	0.952	0.950	0.025	0.950	0.028	0.950	0.028
β_{11}	-0.087	-0.086	0.015	-0.086	0.016	-0.086	0.016
β_{20}	-2.369	-2.316	0.323	-2.377	0.455	-2.401	0.533
β_{21}	0.966	0.950	0.175	0.975	0.229	0.987	0.255
σ_{b1}	0.086	0.082	0.014	0.082	0.013	0.082	0.014
σ_{b2}	0.822	0.721	0.197	0.761	0.197	0.766	0.293
ρ_b	-0.608	-0.637	0.185	-0.788	0.130	–	–
σ_{e1}	0.075	0.076	0.005	0.076	0.005	0.076	0.005
ρ_e	-0.804	-0.767	0.058	–	–	–	–

Table 5.12. Results from simulation study: standard deviations of estimates and of estimated standard errors for small sample and strong correlation.

Parameter	Full model		Reduced model 1		Reduced model 2	
	Est.	SE	Est.	SE	Est	SE
β_{10}	0.026	0.008	0.026	0.004	0.026	0.004
β_{11}	0.016	0.005	0.016	0.002	0.016	0.003
β_{20}	0.442	0.169	0.482	0.128	0.561	0.196
β_{21}	0.213	0.073	0.229	0.056	0.256	0.083
σ_{b1}	0.015	0.002	0.015	0.007	0.015	0.002
σ_{b2}	0.216	0.103	0.223	0.117	0.272	0.075
ρ_b	0.235	0.132	0.219	0.147	–	–
σ_{e1}	0.004	0.002	0.004	0.0003	0.004	0.0002
ρ_e	0.093	0.061	–	–	–	–

Table 5.13. Results from simulation study: average estimates and average estimated standard errors for medium sample and weak correlation.

Parameter	True	Full model		Reduced model 1		Reduced model 2	
	value	Est.	SE	Est.	SE	Est.	SE
β_{10}	0.952	0.953	0.025	0.953	0.024	0.953	0.024
β_{11}	-0.087	-0.088	0.014	-0.088	0.014	-0.088	0.014
β_{20}	-2.369	-2.550	0.429	-2.532	0.467	-2.569	0.501
β_{21}	0.966	1.049	0.215	1.042	0.226	1.056	1.239
σ_{b1}	0.086	0.084	0.012	0.084	0.012	0.084	0.012
σ_{b2}	0.822	0.821	0.213	0.813	0.244	0.816	0.274
ρ_b	-0.608	-0.605	0.181	-0.657	0.193	–	–
σ_{e1}	0.075	0.074	0.004	0.074	0.004	0.074	0.004
ρ_e	-0.201	-0.231	0.135	–	–	–	–

Table 5.14. Results from simulation study: standard deviations of estimates and of estimated standard errors for medium sample and weak correlation.

Parameter	Full model		Reduced model 1		Reduced model 2	
	Est.	SE	Est.	SE	Est	SE
β_{10}	0.022	0.003	0.022	0.003	0.022	0.003
β_{11}	0.016	0.002	0.016	0.002	0.016	0.002
β_{20}	0.620	0.111	0.609	0.232	0.624	0.214
β_{21}	0.291	0.055	0.286	0.094	0.292	0.088
σ_{b1}	0.011	0.001	0.011	0.001	0.011	0.001
σ_{b2}	0.282	0.067	0.280	0.189	0.302	0.080
ρ_b	0.195	0.087	0.194	0.238	–	–
σ_{e1}	0.005	0.0003	0.005	0.0003	0.005	0.0003
ρ_e	0.121	0.014	–	–	–	–

Table 5.15. Results from simulation study: average estimates and average estimated standard errors for medium sample and strong correlation.

Parameter	True	Full model		Reduced model 1		Reduced model 2	
	value	Est.	SE	Est.	SE	Est.	SE
β_{10}	0.952	0.951	0.023	0.951	0.024	0.951	0.024
β_{11}	-0.087	-0.085	0.013	-0.085	0.014	-0.085	0.014
β_{20}	-2.369	-2.468	0.315	-2.537	0.412	-2.537	0.481
β_{21}	0.966	0.973	0.161	1.002	0.201	1.004	0.226
σ_{b1}	0.086	0.082	0.011	0.082	0.010	0.082	0.012
σ_{b2}	0.822	0.733	0.192	0.781	0.176	0.766	0.264
ρ_b	-0.608	-0.627	0.174	-0.768	0.117	–	–
σ_{e1}	0.075	0.075	0.004	0.075	0.004	0.075	0.004
ρ_e	-0.804	-0.767	0.053	–	–	–	–

Table 5.16. Results from simulation study: standard deviations of estimates and of estimated standard errors for medium sample and strong correlation.

Parameter	Full model		Reduced model 1		Reduced model 2	
	Est.	SE	Est.	SE	Est	SE
β_{10}	0.028	0.009	0.028	0.003	0.028	0.003
β_{11}	0.016	0.005	0.016	0.002	0.016	0.002
β_{20}	0.464	0.147	0.500	0.159	0.507	0.156
β_{21}	0.229	0.068	0.244	0.068	0.243	0.065
σ_{b1}	0.011	0.002	0.011	0.005	0.011	0.001
σ_{b2}	0.211	0.092	0.216	0.109	0.238	0.056
ρ_b	0.222	0.105	0.095	0.105	–	–
σ_{e1}	0.004	0.0009	0.004	0.0003	0.004	0.0002
ρ_e	0.066	0.048	–	–	–	–

Then the $(r+1)^{st}$ E-step involves finding

$$Q(\mu,\sigma|\hat{\mu}^{(r)},\hat{\sigma}^{(r)}) = E[l_c|y_i;\hat{\mu}^{(r)},\hat{\sigma}^{(r)}] = -\frac{n}{2}log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} E[(y_i^* - \mu)^2|y_i;\hat{\mu}^{(r)},\hat{\sigma}^{(r)}].$$

From the formulae for the moments of the truncated normal distribution in Johnson et al. (1994), pp.156-157,

$$\begin{aligned}
E[y_i^*|y_i > 0,\hat{\mu}^{(r)},\hat{\sigma}^{(r)}] &= \hat{\mu}^{(r)} + \frac{\phi(\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}})}{\Phi(\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}})}\hat{\sigma}^{(r)} \\
E[y_i^*|y_i < 0,\hat{\mu}^{(r)},\hat{\sigma}^{(r)}] &= \hat{\mu}^{(r)} - \frac{\phi(-\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}})}{\Phi(-\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}})}\hat{\sigma}^{(r)} \\
Var(y_i^*|y_i > 0,\hat{\mu}^{(r)},\hat{\sigma}^{(r)}) &= \hat{\sigma}^{2(r)}(1 - \frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}}\frac{\phi(\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}})}{\Phi(\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}})} - [\frac{\phi(\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}})}{\Phi(\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}})}]^2) \\
Var(y_i^*|y_i < 0,\hat{\mu}^{(r)},\hat{\sigma}^{(r)}) &= \hat{\sigma}^{2(r)}(1 + \frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}}\frac{\phi(-\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}})}{\Phi(-\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}})} - [\frac{\phi(-\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}})}{\Phi(-\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}})}]^2).
\end{aligned}$$

Now using the equalities $\phi(x) = \phi(-x)$, $\Phi(-x) = 1 - \Phi(x)$ and $E\{(y_i^* - \mu)^2|y_i\} = Var(y_i^*|y_i) + (E(y_i^*|y_i) - \mu)^2$, and simplifying the notation by using $\hat{\phi}^{(r)} = \phi(\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}})$ and $\hat{\Phi}^{(r)} = \Phi(\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}})$

$$\begin{aligned}
E\{(y_i^* - \mu)^2|y_i > 0,\hat{\mu}^{(r)},\hat{\sigma}^{(r)}\} &= \hat{\sigma}^{2(r)} + (\hat{\mu}^{(r)} - \mu)^2 - (2\mu - \hat{\mu}^{(r)})\hat{\sigma}^{(r)}\frac{\hat{\phi}^{(r)}}{\hat{\Phi}^{(r)}} \\
E\{(y_i^* - \mu)^2|y_i < 0,\hat{\mu}^{(r)},\hat{\sigma}^{(r)}\} &= \hat{\sigma}^{2(r)} + (\hat{\mu}^{(r)} - \mu)^2 + (2\mu - \hat{\mu}^{(r)})\hat{\sigma}^{(r)}\frac{\hat{\Phi}^{(r)}}{1 - \hat{\Phi}^{(r)}}.
\end{aligned}$$

Therefore $E[l_c|y_i;\hat{\mu}^{(r)},\hat{\sigma}^{(r)}] =$

$$-\frac{n}{2}log(\sigma^2) - \frac{n}{2\sigma^2}[\hat{\sigma}^{2(r)} + (\hat{\mu}^{(r)} - \mu)^2 - (2\mu - \hat{\mu}^{(r)})\hat{\sigma}^{(r)}\frac{\hat{\phi}^{(r)}}{\hat{\Phi}^{(r)}(1 - \hat{\Phi}^{(r)})}(\hat{\Phi}^{(r)} - \frac{n_i}{n})].$$

The $(r+1)^{st}$ M-step of the EM algorithm then reduces to finding the maximum of the above likelihood with respect to μ and σ. From

$$\frac{\partial E(l_c|y_i;\hat{\mu}^{(r)},\hat{\sigma}^{(r)})}{\partial \mu} = \frac{n}{\sigma^2}(\hat{\mu}^{(r)} - \mu + \hat{\sigma}^{(r)}\frac{\hat{\phi}^{(r)}}{\hat{\Phi}^{(r)}(1-\hat{\Phi}^{(r)})}(\hat{\Phi}^{(r)} - \frac{n_i}{n})) = 0$$

we obtain

$$\hat{\mu}^{(r+1)} = \hat{\mu}^{(r)} - \hat{\sigma}^{(r)}\frac{\hat{\phi}^{(r)}}{\hat{\Phi}^{(r)}(1-\hat{\Phi}^{(r)})}(\hat{\Phi}^{(r)} - \frac{n_i}{n}). \tag{5.20}$$

And similarly from $\frac{\partial E(l_c|y_i;\hat{\mu}^{(r)},\hat{\sigma}^{(r)})}{\partial \sigma^2} = 0$ we obtain

$$\hat{\sigma}^{2(r+1)} = \hat{\sigma}^{2(r)}\{1+\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}}\frac{\hat{\phi}^{(r)}}{\hat{\Phi}^{(r)}(1-\hat{\Phi}^{(r)})}(\hat{\Phi}^{(r)}-\frac{n_i}{n})-[\frac{\hat{\phi}^{(r)}}{\hat{\Phi}^{(r)}(1-\hat{\Phi}^{(r)})}(\hat{\Phi}^{(r)}-\frac{n_i}{n})]^2\}. \tag{5.21}$$

Equations (5.20) and (5.21) define the iterative algorithm for finding the parameter estimates of μ and σ^2. Notice that if $\hat{\Phi}^{(r)} = \frac{n_i}{n}$ then $\hat{\mu}^{(r+1)} = \hat{\mu}^{(r)}$ and $\hat{\sigma}^{(r+1)} = \hat{\sigma}^{(r)}$, and hence $\Phi^{-1}(\frac{n_i}{n})$ (the maximum likelihood estimate of $\frac{\mu}{\sigma}$) is the only stationary point for both μ and σ in the EM algorithm. In fact it is the only stationary point also for the ratio $\frac{\mu}{\sigma}$. To prove that define

$$\begin{aligned} g(x;c) &= \frac{\phi(x)(\Phi(x)-c)}{x\Phi(x)(1-\Phi(x))} \\ f_1(x;c) &= 1 - g(x;c) \\ f_2(x;c) &= 1 + x^2 g(x;c) - x^2[g(x;c)]^2. \end{aligned}$$

Notice that $\hat{\mu}^{(r+1)} = \hat{\mu}^{(r)} f_1(\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}};\frac{n_i}{n})$ and $\hat{\sigma}^{(r+1)} = \hat{\sigma}^{(r)}\sqrt{f_2(\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}};\frac{n_i}{n})}$ for $f_2(\frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}};\frac{n_i}{n}) > 0$. Suppose that there is another stationary point $x_s \neq \Phi^{-1}(\frac{n_i}{n})$ for the ratio $\frac{\mu}{\sigma}$. Then $g(x_s;c) \neq 0$, $f_1(x_s;c)^2 = f_2(x_s;c)$ and $f_1(x_s;c)$ must be positive. The equation

$$(1-g(x_s;c))^2 = 1 + x_s^2 g(x_s;c) - x_s^2[g(x_s;c)]^2$$

has the following non-zero solution

$$g(x_s; c) = \frac{x_s^2 + 2}{x_s^2 + 1} > 1,$$

which implies $f_1(x_s; c) < 0$ ($f_2(x_s; c) > 0$) and therefore there is no other stationary point for the ratio $\frac{\mu}{\sigma}$. Notice that the functions above are not defined at $x = 0$, but this is not a problem since $\frac{\mu}{\sigma} = 0$ cannot be a stationary point unless $\Phi(\frac{n_1}{n}) = 0$, and hence we can assume that $x \neq 0$ for the definitions above. The finding that there are no other stationary points except the maximum likelihood estimate is important because otherwise it is not guaranteed that the algorithm will converge to the maximum likelihood estimate of the identifiable ratio.

By the basic property of the EM algorithm, the observed data log-likelihood is non-decreasing at each step of the algorithhm (Dempster, Laird and Rubin (1977), Wu (1983)) and this is true regardless of whether all parameters are identifiable from the observed data or not. Let $l_o(\frac{\mu}{\sigma}) = l_o(\mu, \sigma)$ denote the observed data log-likelihood as a function of the parameters. Then $l_o(\mu, \sigma) = l_c(\mu, \sigma) - l_i(\mu, \sigma)$, where $l_i(\mu, \sigma) = lnf(\mathbf{y}_i^* | \mathbf{y}_i, \mu, \sigma)$. Taking conditional expectations with respect to the observed data

$$l_o(\mu, \sigma) = Q(\mu, \sigma; \mu^*, \sigma^*) - H(\mu, \sigma; \mu^*, \sigma^*),$$

where $Q(\mu, \sigma | \mu^*, \sigma^*) = E\{l_c(\mu, \sigma) | \mathbf{y}, \mu^*, \sigma^*\}$ and
$H(\mu, \sigma | \mu^*, \sigma^*) = E\{l_i(\mu, \sigma)) | \mathbf{y}, \mu^*, \sigma^*\}$.

$$Q(\hat{\mu}^{(r+1)}, \hat{\sigma}^{(r+1)} | \hat{\mu}^{(r)}, \hat{\sigma}^{(r)}) \geq Q(\hat{\mu}^{(r)}, \hat{\sigma}^{(r)} | \hat{\mu}^{(r)}, \hat{\sigma}^{(r)})$$

and

$$H(\hat{\mu}^{(r)}, \hat{\sigma}^{(r)} | \hat{\mu}^{(r)}, \hat{\sigma}^{(r)}) \geq H(\mu^*, \sigma^* | \hat{\mu}^{(r)}, \hat{\sigma}^{(r)})$$

for any μ^*, σ^* in the parameter space by a consequence to the Jensen's inequality (Dempster, Laird and Rubin (1977)). Therefore

$$l_o(\hat{\mu}^{(r+1)}, \hat{\sigma}^{(r+1)}) \geq l_o(\hat{\mu}^{(r)}, \hat{\sigma}^{(r)}).$$

Because the observed data log-likelihood is concave in the parameter $\frac{\mu}{\sigma}$, for any $(\hat{\mu}^{(0)}, \hat{\sigma}^{(0)})$ inside the parameter space the sequence $\{l_o(\hat{\mu}^{(r)}, \hat{\sigma}^{(r)})\}$ is bounded from above and hence will converge to some l^*. If we can assure that the log-likelihood l_o is strictly increasing for all $\frac{\mu}{\sigma} \neq \Phi^{-1}(\frac{n_i}{n})$ then $l^* = l_o(\Phi^{-1}(\frac{n_i}{n}))$. But the likelihood will be strictly increasing if $\frac{\hat{\mu}^{(r+1)}}{\hat{\sigma}^{(r+1)}} \neq \frac{\hat{\mu}^{(r)}}{\hat{\sigma}^{(r)}}$, that is if there are no other stationary points for the ratio $\frac{\mu}{\sigma}$ in the EM-algorithm.

In more complicated cases when there are no closed form expressions for either the E-step or the M-step of the EM algorithm, it is not possible to algebraically verify that there are no other stationary points of the algorithm except those corresponding to stationary points for the observed data log-likelihood. But a modification may be implemented which according to Liu, Rubin and Wu (1999) will avoid that problem. In the context of the example above the modification will amount to setting $\sigma_{\epsilon 2}$ to be equal to its null value at each step of the algorithm rather than using the current estimate. Still, the general question of convergence of the algorithm as initially proposed is of significant interest and justifies further research.

5.7 Model Extensions

The correlated probit model can be generalized to incorporate any combination of binary, continuous or continuous censored data. It can also accommodate ordinal data for which the cutoff points for the underlying continuous random variable are known.

To demonstrate how the extensions proposed above can be carried out we consider the underlying linear mixed model defined in Section 1 in equations (5.2)-(5.5) and denote the first variable by y^*_{i1j} rather than by y_{i1j}. We may observe either $(y^*_{i1j}, y^*_{i2j})^T$, or $I\{y^*_{i1j} > 0\}$, $I\{y^*_{i2j} > 0\}$, or $I\{y^*_{i1j} > \tau_{11}\}$, $I\{y^*_{i1j} > \tau_{12}\}$, ... $I\{y^*_{i1j} > \tau_{1,p_1}\}$, $I\{y^*_{i2j} > \tau_{21}\}$, $I\{y^*_{i2j} > \tau_{22}\}$, ... $I\{y^*_{i2j} > \tau_{2,p_2}\}$, where $\tau_{11}, ...\tau_{1,p_1}, \tau_{21}, ..., \tau_{2,p_2}$ are

known, or

$$y^c_{i1j} = \begin{cases} y^*_{i1j} & if \quad y^*_{i1j} > \gamma_1 \\ \gamma_{y1} & if \quad y^*_{i1j} \leq \gamma_1 \end{cases}$$

$$y^c_{i2j} = \begin{cases} y^*_{i2j} & if \quad y^*_{i2j} > \gamma_2 \\ \gamma_{y2} & if \quad y^*_{i2j} \leq \gamma_2 \end{cases}$$

where γ_1, γ_2, γ_{y1}, γ_{y2} are known, or any combination of the above.

For all those cases the complete data MLE will be the MLE for $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\Sigma}_e$ from the mixed model in Section 5.2. Therefore, the M-step in the EM algorithm will be the same. At each E-step expressions or approximations of $E(\mathbf{y}^*_i|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)})$ and $E(\mathbf{y}^*_i\mathbf{y}^{*T}_i|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)})$ are needed, where $\mathbf{y}_i$ is the observed response vector for subject i. If $\mathbf{y}_{i1} = \mathbf{y}^*_{i1}$ then we don't need to generate values for this response and we just use $E(\mathbf{y}^*_{i1}|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)}) = \mathbf{y}_{i1}$ and $E(\mathbf{y}^*_{i1}\mathbf{y}^{*T}_{i1}|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)}) = \mathbf{y}_{i1}\mathbf{y}^T_{i1}$, and similarly for the other response. Otherwise we do need to generate values.

The binary case has already been described in detail earlier in this chapter. The ordinal case is handled in the same way with the only difference that generated values for the truncated multivariate normal distribution must fall in the region specified by the ordinary response. The censored response is kept as it is if it corresponds to uncensored observation, that is $E(y^*_{i1j}|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)}) = y^c_{i1j}$ and $E(y^*_{i1j}{y^*_{i1j}}^T|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)}) = y^c_{i1j}{y^c_{i1j}}^T$, if $y^c_{i1j} \neq \gamma_{y1}$. If, however, $y^c_{i1j} = \gamma_{y1}$, then values are generated from the truncated normal distribution of $y^*_{i1j}|y^{*(-)}_i, y_{i1j} = \gamma_{y1}$ as in the binary case. (The truncation in this particular example is from above at γ_1.) Then $E(y^*_{i1j}|\mathbf{y}_i, \hat{\boldsymbol{\psi}}^{(r)}) = \frac{1}{m}\sum_{k=1}^m y^{*(k)}_{i1j}$ and the variance is handled similarly.

Another possible extension of the correlated probit model is that more general correlation structures may be incorporated both at the random effects and at the random error level. One example is autoregressive structure, which will allow one to model a variety of longitudinal data sets.

5.8 Future Research

There are several unanswered questions concerning the model and the algorithms proposed in this chapter which justify further research. Establishing convergence for the type of Markovian dependence implied by Liao's Gibbs sampler is certainly one of them. Gu and Kong (1998) propose a stochastic approximation algorithm with the Markov chain Monte Carlo method for incomplete data estimation problems that uses basically the same type of idea. There are also some general results concerning stochastic algorithms with Markovian perturbations in the probability theory literature. These can serve as a basis for establishing convergence properties.

Another interesting question is the choice of weights to assure faster convergence of the Monte Carlo EM algorithm. We noticed that when the initial estimates were far from the maximum likelihood estimates the algorithms had convergence problems, and this can probably be remedied if the weights are wisely chosen. A general result for the identifiability problem as outlined in Section 5.6. is needed for completeness, and this can probably be achieved by extending some of the theorems by Wu (1983).

The problem of checking the assumptions for the correlated probit model is also very important, as misspecification of the model can lead to inconsistent parameter estimates (White, 1982). The effects of incorrect specification of the random effects distribution are of special interest because the problem of joint versus separate fitting of the response variables can be addressed from that perspective. Neuhaus, Hauck and Kalbfleisch (1992) investigated the effects of misspecification of the random effects distribution in mixed-effects logistic models and found that the bias in the parameter estimates is usually small. It will not be surprising if this is the case for the correlated probit model but the issue must be addressed in greater detail.

CHAPTER 6
CONCLUSIONS

6.1 Summary

The goal of this dissertation was to propose and investigate random effects models for repeated measures situations when there are two or more response variables. Our emphasis was on maximum likelihood estimation and on applications with outcomes of different types. We proposed a multivariate generalized linear mixed model that can accommodate any combination of responses in the exponential family. We also considered a correlated probit model that is suitable for mixtures of binary, continuous, censored continuous and ordinal outcomes. Although more limited in area of applicability, the correlated probit model allows for more general correlation structure between the response variables than the corresponding multivariate generalized linear mixed model.

We used two 'real-life' applications for illustration: a developmental toxicity study in mice and a myoelectric activity study in ponies. The first data set had a binary and a continuous response which made it suitable for illustration of both models, and the second data set had a count and a duration response which were fitted as negative binomial and gamma variates in a multivariate generalized linear mixed model. The two data sets were also different in that the mice data had a relatively large number of subjects (about 100) and between 1 and 16 observations per subject, while the pony data had only 6 subjects but many observations per subject. Because the large sample asymptotic theory holds when the number of subjects goes to infinity we used the mice data for illustration of hypothesis testing in Chapter 4.

In Chapter 2 we defined the multivariate generalized linear mixed model by specifying a separate GLMM for each response variable and then combining the models by imposing a joint multivariate normal distribution on the subject-specific random effects. The responses on the same subject were assumed to be conditionally independent given the random effects, which allowed estimation procedures for GLMM to be directly modified for this more general model.

In Chapter 3 we extended three approximate maximum likelihood estimation methods from the univariate to the multivariate generalized linear mixed model. The three methods were Gauss-Hermite quadrature, the Monte Carlo EM algorithm proposed by Booth and Hobert (1999) and the pseudo-likelihood approach proposed by Wolfinger and O'Connell (1993). In addition to parameter estimation we considered approximations of the standard errors based on numerical derivatives in Gauss-Hermite quadrature, on Louis' approximation to the observed information matrix in the Monte Carlo EM algorithm, and on linear mixed model theory in the pseudo-likelihood method. We used a simulated data example and the two 'real-life' data sets for illustration.

Some findings were as follows. The pseudo-likelihood method led to underestimation of some of the parameters and their standard errors for the Bernoulli response. The Gauss-Hermite and the Monte Carlo EM methods performed well but were computationally intensive. The Monte Carlo standard error estimates showed high level of variability. In an attempt to counter that effect we proposed to pool the estimates of the information matrix from the last three iterations of the algorithm. This seemed to alleviate but did not solve the problem. In neither of the three examples there was evidence of significant efficiency gains from fitting the responses together rather than separately. This issue was further studied with a simulation study in Chapter 5. Because of the extra-Poisson variability present in the count outcome of the pony

data we had to apply a special nested EM algorithm to fit the negative-binomial distribution for the count response.

In Chapter 4 we considered hypothesis testing in the multivariate GLMM. Under the assumption that the regularity conditions are satisfied, asymptotic likelihood theory can be applied to form hypothesis tests and confidence intervals concerning the parameters of interest. We suggested to test the significance of the fixed effects using approximations to the Wald and to the likelihood ratio statistics and to estimate the random effects using approximations to the conditional mean $E(\mathbf{b}_i|\mathbf{y}_i)$. We extended a global variance component score test originally proposed by Lin (1997) for the univariate GLMM to the multivariate GLMM, and outlined an approach to approximate the score statistics for subsets of the variance components using Gauss-Hermite or Monte Carlo approximations. We also proposed a score test for checking the conditional independence assumption between the response variables and developed Gaussian quadrature and Monte Carlo approximations for the case of one binary and one continuous response. Because we could use all three statistics (the Wald, the score and the likelihood ratio) to test for conditional independence we compared the performance of the approximations on checking the conditional independence assumption in the developmental toxicity example. We also designed a small simulation study and observed that the Gaussian quadrature approximations to the three statistics led to almost identical results although usually the score statistic had the largest and the likelihood ratio statistic had the smallest value. The Monte Carlo approximation to the likelihood ratio statistic was not entirely consistent with the other approximations and may require larger simulation sample size.

In Chapter 5 we introduced the correlated probit model as an alternative to the multivariate GLMM when conditional independence did not hold. This model was first considered by Catalano and Ryan (1992) who marginalized it and used GEE

methods to obtain estimates. We developed a Monte Carlo EM algorithm for maximum likelihood estimation which could be regarded as an extension to an approach proposed by Chan and Kuk (1997) for binary data. We applied the method to the developmental toxicity example. Because of the computational inefficiency of the algorithm we considered a modification based on stochastic approximations (Liao (1999), Lavielle et al. (1999)) which led to a significant decrease in the time for model fitting. To address the issue of advantages of joint over separate analyses of the response variables we designed a simulation study to investigate possible efficiency gains in a multivariate analysis. Noticeable increase in the estimated standard errors was observed only in the binary response case for small number of subjects and observations per subject and for high correlation between the outcomes. We also briefly considered an identifiability issue for one of the variance components.

In conclusion, the proposed models are appropriate for multivariate repeated measures applications when subject specific inference is of main interest. They allow one to answer intrinsically multivariate questions such as estimation of the probability of malformation and/or low fetal weight at any given dose of ethylene glycol in the mice example. Multivariate analysis also allows to maintain the proper significance levels in hypothesis tests concerning several outcome variables. However, efficiency gains concerning individual parameters are noticeable only for small samples and highly correlated outcomes. Therefore, if multivariate inference is not the emphasis of the analysis, separate analyses of the outcome variables may be preferrable because of the availability of software for univariate repeated measures such as SAS PROC NLMIXED.

6.2 Future Research

There are variety of interesting topics for further research concerning the models proposed in this dissertation. One of them is a comparison of the numerical and

stochastic approximations proposed here to the analytical approximations proposed by Lin (1997) for the score statistic for variance components. It is also of interest to develop methods to assess error in Gaussian quadrature and Monte Carlo approximations of the test statistics. This way the number of quadrature points and the Monte Carlo sample size can be chosen appropriately to achieve certain precision. A better approach to dealing with the variability of Monte Carlo standard errors than just pooling several estimates of the information matrix is also needed. The good performance of the stochastic approximation to the observed information matrix proposed by Lavielle et al. (1999) is an indication that if this idea could be incorporated in the Monte Carlo EM algorithm more stable standard error estimates could possibly be achieved.

We addressed the issue of efficiency gains of joint over separate fitting of the response variables via a simulation study in Chapter 5. We restricted our attention to the case of one binary and one continuous response and to the particular structure of the developmental toxicity example so our findings have limited applicability. Hence it is justified to design additional simulation studies to investigate different settings. We also did not address the issue of bias of the maximum likelihood estimates when the model was misspecified. The maximum likelihood estimates are guaranteed to be consistent only under the assumption that the model is chosen correctly. Hence if the responses are fitted separately while the true model requires them to be fitted jointly the regression parameter estimates may show some bias. It is of special interest to determine whether this happens and if it does what the magnitude of the bias is.

A variety of other questions are also not yet resolved. More research is needed on improving the computational methods for obtaining maximum likelihood estimates, investigating the performance of the asymptotic tests when small samples are present, assessing goodness-of-fit and performing variable and model selection, studying the effect of departure from the parametric assumptions on the estimates, developing

methods for residual diagnostics and outlier detection. One of the most important challenges, however, is to develop reliable, fast and user-friendly software for generalized linear mixed models and their extensions. PROC NLMIXED in SAS is a step in this direction but further extensions will be needed. Without widely available software those models remain an interesting new development without much practical applicability.

REFERENCES

Abramowitz, M., & Stegun, I. (1972). *Handbook of Mathematical Functions.* New York: Dover.

Agresti, A. (1990). *Categorical Data Analysis.* New York: John Wiley.

Agresti, A. (1997). A model for repeated measurements of a multivariate binary response. *Journal of the American Statistical Association, 92,* 315–321.

Aitchison, J., & Ho, C. H. (1989). The multivariate Poisson-log normal distribution. *Biometrika, 76,* 643–653.

Aitchison, J., & Silvey, S. D. (1957). The generalization of probit analysis to the case of multiple responses. *Biometrika, 44,* 131–140.

Ashford, J. R., & Sowden, R. R. (1970). Multi-variate probit analysis. *Biometrics, 26,* 535–546.

Bera, A. K., & Bilias, Y. (1999). Rao's score, Neyman's C(α) and Silvey's LM tests: An essay on historical developments and some new results. *To appear in Journal of Statistical Planning and Inference.*

Blackwell, B., & Catalano, P. J. (1999a). Correlated random effects latent variable models for multivariate ordinal repeated measures bioassays. *Unpublished manuscript.*

Blackwell, B., & Catalano, P. J. (1999b). A random effects latent variable model for ordinal data. *Unpublished manuscript.*

Bliss, C. I. (1934a). The method of probits. *Science, 79,* 38–39.

Bliss, C. I. (1934b). The method of probits – a correction. *Science, 79,* 409–410.

Bliss, C. I. (1935a). The calculation of the dosage-mortality curve. *Annals of Applied Biology, 22,* 307–333.

Bliss, C. I. (1935b). The comparison of dosage-mortality data. *Annals of Applied Biology, 22,* 307–333.

Booth, J. G., & Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association, 93,* 262–272.

Booth, J. G., & Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological, 61*, 265–285.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association, 88*, 9–25.

Buhler, W. J., & Puri, P. S. (1966). On optimal asymptotic tests of composite hypotheses with several constraints. *Z. Wahracheinlichkeitstheorie verw, 5*, 71–88.

Catalano, P. J. (1994). Bivariate modeling of clustered continuous and ordered categorical outcomes. *Statistics in Medicine, 16*, 883–900.

Catalano, P. J., & Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association, 87*, 651–658.

Chan, J. S. K., & Kuk, A. Y. C. (1997). Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics, 53*, 86–97.

Chant, D. (1974). On asymptotic tests of complete hypotheses in nonstandard conditions. *Biometrika, 61*, 291–298.

Coull, B. (1997). Subject-specific modelling of capture-recapture experiments. *Ph.D. dissertation, University of Florida, Gainesville, Dept. of Statistics.*

Cox, D. R. (1972). The analysis of multivariate binary data. *Applied Statistics, 21*, 113–120.

Cox, D. R., & Wermuth, N. (1992). Response models for mixed binary and quantitative variables. *Biometrika, 79*, 441–461.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (c/r: p22-37). *Journal of the Royal Statistical Society, Series B, Methodological, 39*, 1–22.

Diggle, P. J., Liang, K.-Y., & Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Oxford: Clarendon Press.

Doornik, J. A. (1998). *Object-oriented Matrix Programming using Ox version 2.0.* Kent: Timberlake Consultants.

Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information (c/r: p482-487). *Biometrika, 65*, 457–481.

Fahrmeir, L., & Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models.* New York: Springer-Verlag.

Finney, D. J. (1964). *Probit analysis.* Cambridge: Cambridge University Press.

Fitzmaurice, G. M., & Laird, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *Journal of the American Statistical Association*, *90*, 845–852.

Foutz, R. V. (1977). On the unique consistent solution to the likelihood equations. *Journal of the American Statistical Association*, *72*, 147–148.

Gaddum, J. H. (1933). Reports on biological standards. III. Methods of biological assay depending on a quantal response. *Spec. Rep. Ser. Med. Res. Coun.*, London, no. 183.

Galecki, A. T. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics, Part A – Theory and Methods*, *23*, 3105–3119.

Gu, M. G., & Kong, F. H. (1998). A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems. *Proceedings of the National Academy of Sciences*, *95*, 7270–7274.

Haber, M. (1986). Testing for pairwise independence. *Biometrics*, *42*, 429–435.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems (c/r: p338-340). *Journal of the American Statistical Association*, *72*, 320–338.

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, *50*, 933–944.

Heitjan, D. F., & Sharma, D. (1997). Modelling repeated-series longitudinal data. *Statistics in Medicine*, *16*, 347–355.

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *The Annals of Mathematical Statistics*, *42*, 1977–1991.

Hobert, J. P., & Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, *91*, 1461–1473.

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, *42*, 805–820.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous Univariate Distributions. Volume 1 (Second Edition)*. New York: Wiley-Interscience.

Lavielle, M., Delyon, B., & Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *To appear in the Annals of Statistics*.

Lee, L. (1993). Multivariate tobit models in econometrics. *Handbook of Statistics Volume 11: Econometrics*, 145–173.

Lee, Y., & Nelder, J. A. (1996). Hierarchical generalized linear models (disc: p656-678). *Journal of the Royal Statistical Society, Series B, Methodological*, *58*, 619–656.

Lefkopoulou, M., Moore, D., & Ryan, L. (1989). The analysis of multiple correlated binary outcomes: Application to rodent teratology experiments. *Journal of the American Statistical Association*, *84*, 810–815.

Lesaffre, E., & Molenberghs, G. (1991). Multivariate probit analysis: A neglected procedure in medical statistics. *Statistics in Medicine*, *10*, 1391–1403.

Lester, G., Merritt, A., Neuwirth, L., Widenhouse, T., Steible, C., & Rice, B. (1998a). Effect of α_2-adrenergic, cholinergic, and nonsteroidal anti-inflammatory drugs on myoelectric activity of ileum, cecum, and right ventral colon, and on cecal amptying of radiolabeled markers in clinically normal ponies. *American Journal of Veterinary Medicine*, *59*, 320–327.

Lester, G., Merritt, A., Neuwirth, L., Widenhouse, T., Steible, C., & Rice, B. (1998b). Effect of erythromycin lactobionate on myoelectric activity of ileum, cecum, and right ventral colon, and cecal emptying of radiolabeled markers in clinically normal ponies. *American Journal of Veterinary Medicine*, *59*, 328–335.

Lester, G., Merritt, A., Neuwirth, L., Widenhouse, T., Steible, C., & Rice, B. (1998c). Myoelectric activity of the ileum, cecum, and right ventral colon, and cecal emptying of radiolabeled markers in clinically normal ponies. *American Journal of Veterinary Medicine*, *59*, 313–319.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

Liang, K.-Y., & Zeger, S. L. (1989). A class of logistic regression models for multivariate binary time series. *Journal of the American Statistical Association*, *84*, 447–451.

Liao, J. (1999). A simplified and accelerated Monte Carlo EM algorithm with application to a hierarchical mixture model. *To appear in Statistica Sinica.*

Lin, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika*, *84*, 309–326.

Lindsey, J. K. (1993). *Models for Repeated Measurements.* Oxford: Clarendon Press.

Liu, C., Rubin, D. B., & Wu, Y. N. (1994). Parameter expansion to accelerate EM - the PX-EM algorithm. *Biometrika*, *81*, 624–629.

Liu, Q., & Pierce, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, *81*, 624–629.

Long, S. (1997). *Regression models for categorical and limited dependent variables.* London: Sage Publications.

Longford, N. T. (1993). *Random Coefficient Models.* Oxford: Oxford University Press.

Lundbye-Christensen, S. (1991). A multivariate growth curve model for pregnancy. *Biometrics, 47*, 637–657.

Matsuyama, Y., & Ohashi, Y. (1997). Mixed models for bivariate response repeated measures data using Gibbs sampling. *Statistics in Medicine, 16*, 1587–1601.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models (Second Edition).* New York: Chapman & Hall.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association, 92*, 162–170.

McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology, 4*, 103–120.

Meng, X.-L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika, 80*, 267–278.

Moran, P. A. P. (1971). Maximum-likelihood estimation in non-standard conditions. *Proceedings of the Cambridge Philosophical Society, 70*, 441–450.

Natarajan, R., & McCulloch, C. E. (1995). A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika, 82*, 639–643.

Neuhaus, J. M., Hauck, W. W., & Kalbfleisch, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika, 79*, 755–762.

Neyman, J. (1959). Optimal asymptotic test of composite statistical hypothesis. *In U. Grenader, Ed., Probability and Statistics, the Harald Cramér volume*, Uppsala: Almqvist and Wiksell, 213–234.

Neyman, J., & Pearson, E. (1928). On the use of interpretation of certain test criteria for purpose of statistical inference. *Biometrika, 20*, 175–240.

Ochi, Y., & Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika, 71*, 531–543.

Olkin, L., & Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *Annals of Mathematical Statistics, 32*, 448–465.

Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M., & Fisher, M. R. (1996). A survey of methods for analyzing clustered binary response data. *International Statistical Review, 64*, 89–118.

Pinheiro, J. C., & Bates, D. M. (1995). Approximation to the loglikelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics, 4*, 12–35.

Price, C. J., Kimmel, C. A., Tyl, R. W., & Marr, M. C. (1985). The developmental toxicity of ethylene glycol in rats and mice. *Toxicological Applications in Pharmacology, 81*, 825–839.

Rao, C. R. (1947). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society, 44*, 50–57.

Regan, M., & Catalano, P. (1999). Likelihood models for clustered binary and continuous outcomes: application to developmental toxicology. *Unpublished manuscript.*

Reinsel, G. (1982). Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *Journal of the American Statistical Association, 77*, 190–195.

Reinsel, G. (1984). Estimation and prediction in a multivariate random effects generalized linear model. *Journal of the American Statistical Association, 79*, 406–414.

Rochon, J. (1996). Analyzing bivariate repeated measures for discrete and continuous outcome variables. *Biometrics, 52*, 740–750.

Rosner, B. (1992). Multivariate methods for clustered binary data with multiple subclasses, with application to binary longitudinal data. *Biometrics, 48*, 721–731.

Sammel, M. D., Ryan, L. M., & Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society, Series B, 59*, 667–678.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika, 78*, 719–727.

Self, S. G., & Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association, 82*, 605–610.

Tanner, M. A. (1991). *Tools for Statistical Inference: Observed Data and Data Augmentation Methods.* New York: Springer-Verlag.

Ten Have, T. R. (1996). A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses. *Biometrics, 52*, 473–491.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica, 26*, 24–36.

van Dyk, D. (1999). Nesting EM algorithms for computational efficiency. *To appear in Statistica Sinica.*

Vonesh, E. F. (1992). Non-linear models for the analysis of longitudinal data (disc: p1955-1963). *Statistics in Medicine, 11*, 1929–1954.

Wald, A. (1943). Tests of statistical hypothesis concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc., 54*, 426–482.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician, 39*, 95–101.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*, 1–26.

Wolfinger, R. (1998). Towards practical application of generalized linear mixed models. *Proceedings of 13th international workshop in statistical modeling, Marx, B. and Friedl, H. (eds.)*, 388–395.

Wolfinger, R., & O'Connell, M. (1993). Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation, 48*, 233–243.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics, 11*, 95–103.

Zeger, S. L., & Karim, M. R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *Journal of the American Statistical Association, 86*, 79–86.

Zeger, S. L., & Liang, K.-Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine, 11*, 1825–1839.

Zeger, S. L., Liang, K.-Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach (corr: V45 p347). *Biometrics, 44*, 1049–1060.

Zeger, S. L., Liang, K.-Y., & Self, S. G. (1985). The analysis of binary longitudinal data with time-independent covariates. *Biometrika, 72*, 31–38.

Zeger, S. L., & Qaqish, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics, 44*, 1019–1031.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association, 57*, 348–368.

Zhao, L. P., Prentice, R. L., & Self, S. G. (1992). Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society, Series B, Methodological, 54*, 805–811.

BIOGRAPHICAL SKETCH

Ralitza Gueorguieva was born in Sofia on April 25, 1971. She graduated from the Mathematical High School of Sofia and simultaneously obtained a correspondence degree from the English Language High School in Sofia. In 1989 Ralitza was accepted in the University of Sofia and enrolled as a student in computer science. Five years later she graduated with a Master of Science degree in computer science and obtained additional certification as a teacher in mathematics and computer science. She also spent one semester as an exchange student in Slippery Rock University in Pennsylvania in the fall of 1991.

Ralitza was accepted as a graduate student in the Department of Statistics at the University of Florida in the Fall of 1994. While at the University of Florida she worked as a teaching and research assistant. She obtained her Master of Statistics degree in August 1996 and then proceeded in the Ph.D. program. After graduating from the University of Florida, Ralitza will spend another year in Gainesville teaching an undergraduate statistics class and working in the Perinatal Data Systems group.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Alan Agresti, Chairman
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

James Booth
Associate Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Randolph Carter
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Malay Ghosh
Distinguished Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Monika Ardelt
Assistant Professor of Sociology

This dissertation was submitted to the Graduate Faculty of the Department of Statistics in the College of Liberal Arts and Sciences and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

December 1999

Dean, Graduate School